Kalsgatos Revista de Filosofia
https://doi.org/10.23845/

TRADUÇÃO[1]

SÃO OS SERES HUMANOS ROBÔS HUMEANOS?

ARE HUMAN BEINGS HUMEAN ROBOTS?[2] – GONZALO GÉNOVA[3] E IGNÁCIO NAVARRO[4]

Ismail FAGUNDES
Doutorando em Filosofia pelo Programa de Pós-graduação em Filosofia da Universidade de Caxias do Sul. Bolsista PROSUC/CAPES modalidade II.
E-mail: ismailfagundes@gmail.com

Mariana Rocha BERNARDI
Doutoranda em Filosofia pelo Programa de Pós-graduação em Filosofia da Universidade de Caxias do Sul. Bolsista PROSUC/CAPES modalidade I.
E-mail: mrocha2@ucs.br

ABSTRACT
David Hume, the Scottish philosopher, conceives reason as the slave of the passions, which implies that human reason has predetermined objectives it cannot question. An essential element of an algorithm running on a computational machine (or Logical Computing Machine, as Alan Turing calls it) is its having a predetermined purpose: an algorithm cannot question its purpose, because it would cease to be an algorithm. Therefore, if self-determination is essential to human intelligence, then human beings are neither *Humean* beings, nor computational machines. We examine also some objections to the Turing Test as a model to understand human intelligence.

KEYWORDS:
Human nature, free will, self-determination, algorithm, computational machine.

RESUMO
David Hume, o filósofo escocês, concebeu a razão como a escrava das paixões, o que implica que a razão humana tenha objetivos predeterminados que não possa questionar. Um elemento essencial de um algoritmo em execução em uma máquina computacional (ou *máquina de computação lógica*,

[1]*O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.*
[2] A publicação original se encontra no link https://www.tandfonline.com/doi/full/10.1080/0952813X.2017.1409279 **Journal of Experimental & Theoretical Artificial Intelligence** 30(1):177–186, January 2018.
[3] Gonzalo Génova é professor associado de Engenharia de Software do Departamento de Engenharia e Ciência da Computação na Universidade Carlos III de Madri. É mestre em Filosofia e doutor em Engenharia e ciência da Computação. Entre seus interesses de pesquisa estão a filosofia da ciência da informação e modelagem de linguagens em engenharia de software.
[4] Ignacio Quintanilla Navarro é doutor em filosofia e psicólogo. Professor do Departamento de Estudos Educativos da Universidade Complutense. Há trinta anos investiga o impacto da tecnologia nas escolas, ideias e argumentações.

como Alan Turing chamava) é que ele tem um propósito predeterminado: Um algoritmo não pode questionar seu propósito, porque assim ele deixaria de ser um algoritmo. Portanto, se a autodeterminação é essencial para a inteligência humana, então seres humanos não são nem *seres Humeanos*[5], nem máquinas computacionais. Examinamos também algumas objeções ao teste de Turing como um modelo para entender a inteligência humana.

PALAVRAS-CHAVE:
Natureza humana, Livre-arbítrio, Autodeterminação, Algoritmo, Máquina computacional.

1. INTRODUÇÃO

Ao longo da história, seres humanos sempre refletiram sobre si mesmos, inicialmente se comparando com algo não humano, por exemplo, divindades, animais ou natureza. A instância que hoje preenche essa função é principalmente das máquinas artificiais e, em particular, das máquinas artificiais inteligentes: os robôs são o mais humano de nossas máquinas. A inteligência artificial desempenha, portanto, um papel-chave na atual compreensão de nós mesmos. Assumimos, então, que perguntando sobre a natureza e os limites da inteligência que nós produzimos hoje é um quadro apropriado para investigar a essência da condição humana.

Neste artigo, queremos mostrar a conexão entre a concepção de Hume sobre a natureza humana e a concepção moderna de robôs tal como eles são na realidade (robôs na ficção científica apresentam problemas diferentes, o que inclusive reforça nossa tese). Mesmo se, possivelmente o conceito de "robô" tivesse se mostrado profundamente estranho para Hume, a verdade é que sua concepção de razão enquanto "a escrava das paixões" antecipou a moderna concepção de máquina computacional: nós chamamos essa sua concepção de um *robô Humeano*[6], ou seja, uma inteligência instrumental a serviço de objetivos predeterminados. De fato, se para nós humanos do século 21, é tentador considerar a nós mesmos robôs biológicos sofisticados, é só porque previamente aceitamos o paradigma *Humeano* do século 18 da razão enquanto escrava das paixões. Estamos

5 Seres *Humeanos*, de acordo com a elaboração dos autores, são os seres cuja razão se submete ao crivo das paixões. Neste resumo, eles pretendem apresentar a situação de que os seres humanos não são seres inclinados a apenas uma possibilidade, ou seja, nem totalmente afeitos à razão escravizada pelas paixões, mas tampouco simples máquinas de computação de dados. A autodeterminação do ser humano, aqui, tal como colocada pelos autores, parece indicar um meio termo entre as ideias, desconsiderando-se qualquer pretenso determinismo.

6 Optamos por destacar os termos *Humeano*(s) e *Humeana*(s), para diferenciar de humano(s) e humana(s), deixando em itálico. No original, os autores optam por escrever a adjetivação de Hume com letra maiúscula, o que mantivemos na tradução.

propensos a acreditar que somos robôs, porque primeiro aceitamos que a razão nem escolhe nem prioriza seus fins. (Isso não significa necessariamente que os fins do comportamento humano são completamente definidos*;* apenas significa que sua seleção e priorização não são *racionais*).

Nosso propósito aqui é desafiar a viabilidade de um programa de pesquisa inspirado na concepção de Hume acerca da natureza humana, e nós achamos que o atual desenvolvimento da inteligência artificial apoia nossa tese.

2. DAVID HUME: A RAZÃO É A ESCRAVA DAS PAIXÕES

David Hume (1711-1776) escreveu em *O Tratado da Natureza Humana*, na seção dedicada aos motivos influenciadores da vontade, que "razão é, e deve somente ser a escrava das paixões, e nunca pode pretender qualquer outra função que não o de servir e obedecer a elas" (HUME, 1739). O plano filosófico de Hume era aplicar às ciências do homem o método da "filosofia experimental" (isto é, filosofia natural, ou física), e estender à filosofia em geral as frutíferas autolimitações metodológicas da física Newtoniana (COPLESTON, 1999) [7]. Hume queria compreender a mente humana assim como Isaac Newton compreendeu o cosmos, através da adoção de uma abordagem mecanicista à inteligência humana. Seres humanos são atraídos por paixões, e se movem em direção a uma paixão concreta que só pode ser resistida com a ajuda de uma paixão mais forte e oposta, da mesma forma que forças físicas operam sobre os corpos. Paixões, neste sentido, devem ser entendidas enquanto vontades ou impulsos, deixando de lado a categoria de emoções. Nesta concepção de natureza humana, o papel da razão é elaborar uma estratégia que melhor preencha o conjunto de paixões; mas a razão nem questiona nem escolhe as paixões que têm que servir. Para além desse preciso significado histórico da afirmação de que a "razão é a escrava das paixões", pensamos que Hume propôs uma sugestiva descrição da razão instrumental que antecipa e prepara um modelo algorítmico moderno de inteligência.

De fato, se nós nos compreendermos como máquinas, isso é quase inevitável – à luz do atual estado da arte em inteligência artificial – para identificar as paixões e impulsos *Humeanos* ou

7 Aparentemente, a autolimitação das metodologias seriam o que o tornaria frutífero, no sentido de inteligível, o conhecimento, a partir da experiência sensível, tal qual a física Newtoniana pretendeu.

orientar a objetivos predeterminados ou estágios finais em nossa 'programação natural', e a razão *Humeana* com cálculos visando à optimização do alcance destes objetivos.

Como iremos argumentar mais tarde, achamos que essa é uma noção bem restrita e discutível de inteligência humana. Contudo, parece interessante inicialmente aceitar a sugestão de Hume a fim de explorar algumas abordagens atuais à relação entre algoritmos e mentes humanas. Se assumirmos, como o próprio Hume fez, que as paixões humanas e os desejos admitem um quase-algoritmo ou uma formalização mecânica – através de um modelo homeostático, por exemplo – podemos permitir um modelo teorético comum para explicar os significados e os fins de todo comportamento humano observável.

Na abordagem sentimentalista de Hume à ética, a felicidade é adquirida quando as paixões são satisfeitas. Paixões são forças de atração as quais, em analogia às forças Newtonianas, podem ser resistidas apenas se há outra paixão atraindo em direção oposta. Desde que existem múltiplas paixões (orgulho e humildade, amor e ódio etc.), e eles apontam em diferentes direções, o comportamento resultante é um equilíbrio de forças que tende a preservar a paz interna e a tranquilidade (*homeostase*). Neste modelo, a razão é entendida primariamente como uma ferramenta de otimização (razão técnica ou instrumental, portanto), usada para calcular o comportamento que melhor satisfaz as paixões envolvidas e que exige menos esforço do sujeito. *A razão é a escrava das paixões*: não questiona aquelas paixões que são irresistivelmente impostas a ela, nem seus objetivos, nem sua força de atração; paixões e objetivos são pré-racionais ou meta-racionais.

Quando existem diversas paixões opostas em disputa, e não é possível satisfazer a todas elas (por exemplo, o desejo de furtar algo e o medo de ser preso), é, no entanto, possível de definir uma função multiobjetivo, uma espécie de média ponderada de satisfações de todos os seus objetivos (ou algum outro tipo de função agregada, não necessariamente média ponderada). Esta função (para ser maximizada) é capaz de unificar em um único objetivo a pluralidade de atrações, e assim ser capaz de trazer ordem às paixões individuais. O ponto chave é que a forma desta função (seus pesos, por assim dizer) é também pré-racional ou meta-racional. Como uma consequência de ser a escrava das paixões, a *razão não impõe nenhuma hierarquia entre objetivos*. A função multiobjetivo a ser maximizada, que expressa a ordem e hierarquia das paixões, é também imposta sobre a razão.

Neste esquema, onde a Razão é integrada ao domínio das paixões como um cálculo algorítmico, a vontade é reduzida a um tipo de motivação psicológica também. O que geralmente chamamos 'vontade' não pode ser nada além de automático: uma vez que o melhor caminho é conhecido, tudo que resta é começar, dar a ordem, mas não propriamente 'decidir'. O que está implícito aqui é uma negação radical da liberdade humana no sentido usual do termo, ao qual nos referiremos mais tarde.

3. INTELIGÊNCIA HUMANA, INTELIGÊNCIA ROBÓTICA

Desde o século XVII, a cultura ocidental desenvolveu um programa epistemológico onde podemos verdadeiramente entender apenas o que somos capazes de replicar ou produzir, até mesmo em condições ideais. Portanto, entender a inteligência natural humana requer, ou pelo menos o entendimento fica facilitado por produzir primeiro a inteligência artificial. Contudo, algumas ressalvas filosóficas devem ser apresentadas antes de prosseguirmos com nossa argumentação:

1. Artificial não significa necessariamente não-orgânico.

2. Inteligência não é necessariamente uma qualidade exclusiva de seres humanos; além disso, talvez a inteligência humana não seja o arquétipo de inteligência.

3. Nós não podemos assumir que inteligência é o elemento chave definidor da condição humana, mesmo de uma perspectiva cognitiva.

4. Nós sabemos e entendemos inteligência artificial muito melhor do que a inteligência humana natural, porque produzimos a primeira, enquanto a última nos foi dada.[8]

5. Pesquisa em inteligência artificial abrange mais aspectos do que o desempenho de algoritmos em uma máquina computacional, tais como: ter emoções, perceber o mundo enquanto uma totalidade, ter consciência de si próprio, ter consciência pessoal, ter desejos próprios, ter a capacidade de escolher entre bem e mal e assim por diante.

8 O termo originalmente utilizado no texto foi "given", o qual optamos por traduzir por "dado", no sentido de que é algo que não foi exercitado, mas recebido naturalmente.

6. Não sabemos exatamente o que significa ser inteligente, nem mesmo no sentido humano restrito; portanto, não sabemos se este tipo de inteligência pode ser adequadamente expressa em termos algorítmicos.

7. Não há uma definição incontestável de algoritmo.

Mesmo estando ciente destas dificuldades, parece inegável que a possibilidade de um artefato inorgânico simular processos mentais que podem ser observados em um ser humano possui uma grande importância teórica, tanto para nossa noção de humanidade quanto para o desenvolvimento da inteligência artificial.

4. ALAN TURING: O QUE É UMA MÁQUINA COMPUTACIONAL

Um robô é geralmente definido como um dispositivo mecânico que é controlado por um computador que está executando um programa. Neste artigo, não estamos preocupados com o aspecto físico do robô, isto é, se lembra ou não um ser humano (androide masculino ou ginóide feminino); nem o fato de o robô poder ser feito de materiais inorgânicos, materiais orgânicos ou uma mistura de ambos. Nós apenas estamos preocupados com o fato de um robô ser controlado por um algoritmo, ou conjunto de algoritmos; um robô é, neste sentido, uma máquina algorítmica ou computacional.

Um algoritmo pode ser preliminarmente definido como um procedimento baseado em regras que obtém um resultado desejado a partir de um número finito de passos. Alan Turing firmou as bases da noção moderna de algoritmo, estabelecendo que um método computacional é *efetivo* (também conhecido como mecânico) se puder ser realizado por uma Máquina de Turing (TURING, 1936), ou, como Turing mesmo chama, uma Máquina Lógica Computacional (TURING, 1948). Esta é a substância da tese de Church-Turing (COPELAND, 2002).

Contudo, e talvez surpreendentemente, há uma falta de consenso satisfatório na definição de algoritmo (VARDI, 2012). Um estudo recente feito por Hill (HILL, 2015) as abordagens existentes à noção de algoritmo, a partir de definições semiformais como a realizada por Donald Knuth, “um algoritmo é uma sequência finita de regras que fornecem uma sequência de operações para resolver um tipo específico de problema (KNUTH, 1997), para os mais formais.

Após algumas análises, Hill oferece uma definição preliminar: “Um algoritmo é uma estrutura de controle composta, finita, abstrata, efetiva, imperativamente dada”. O autor

argumenta que esta definição é incompleta porque não leva em conta *o que o algoritmo na realidade faz*: "Há mais em um algoritmo do que seu procedimento". Portanto, a definição é mais refinada conforme segue (sua adição em itálico): "Um algoritmo é uma estrutura de controle composta, finita, abstrata, efetiva, imperativamente dada, *realizando um determinado propósito sob determinadas disposições*.". A razão que ela dá para esta adição – com a qual concordamos completamente – é a necessidade de manifestar a *intencionalidade* dos algoritmos. Um algoritmo não é algo que simplesmente acontece, mas alguma coisa que acontece *com um propósito*, por exemplo, fazer algo *para alguém*.

Este senso de utilidade ou de propósito é compartilhado por máquinas em geral, em contraste com outros tipos de artefatos que podem não ter um propósito útil, como obras de arte. Como tem sido extensivamente lidado pela filosofia da tecnologia (KROES, 2010), uma máquina tem uma dupla natureza que abrange tanto sua *estrutura* física quanto a *função* que deve realizar. É o sucesso ou o fracasso em cumprir sua função que permite que nós digamos se a máquina funciona apropriadamente ou não. Portanto, *uma máquina não pode ser definida e contabilizada como tal sem a referência ao seu propósito*.

Tomemos, por exemplo, uma máquina de jogo, projetada para jogar contra um humano. Inicialmente, a máquina tem o objetivo de ganhar o jogo. Se o jogo é muito simples (como o jogo da velha), projetando uma estratégia (um algoritmo) para ganhar, ou ao menos para não perder, é bastante fácil. No caso do xadrez, a complexidade do jogo não permitiu, até agora, uma estratégia, embora, com a tecnologia atual, a maioria dos jogadores humanos perderia contra um jogador artificial bastante comum de xadrez.

Um tipo um pouco diferente de máquina de xadrez pode incluir um certo grau de aleatoriedade em suas 'decisões', ou pode ser capaz de autolimitar a eficácia de suas estratégias para configurar um nível de dificuldade acessível, para que o jogador humano ainda aproveite o jogo e não desista cedo demais.

Estes dois tipos de máquinas de xadrez têm objetivos sensivelmente diferentes: ou ganhar o jogo, ou então deixar o jogador humano aprender como jogar melhor e desfrutar do processo de aprendizagem. Mesmo assim, em cada caso a máquina tem um propósito predeterminado ou função que o define. O que não esperamos de uma máquina de xadrez do primeiro tipo (por exemplo, desenvolvida para ganhar) é que escolha perder o jogo... *Pode falhar em atingir seu objetivo, mas não pode mudar seu objetivo*.

É claro, podem existir diferentes níveis de seleção de objetivos. Existem na realidade algoritmos que podem dinamicamente alterar seus objetivos, priorizá-los, etc. Então eles são capazes de executar algum tipo de meta-raciocínio em relação aos objetivos a serem alcançados. Contudo, esses algoritmos de dinâmicos de seleção de objetivos, por sua vez, não analisam a si próprios e mudam seus objetivos. Eles estão na verdade obedecendo objetivos de ordem superior (meta-objetivos) para selecionar subobjetivos convenientes. Eles não podem decidir parar de se comportar como algoritmos de seleção de objetivos. Portanto, essa objeção não afeta nosso argumento.

Resumindo, um elemento essencial de um algoritmo executado em uma máquina computacional é predeterminado por seu propósito: *um algoritmo não pode questionar seu propósito, porque deixaria de ser um algoritmo*. Portanto, em um certo sentido, um robô algorítmico é o escravo inteligente de seus propósitos... é um *robô Humeano*. Por outro lado, se a razão humana pode questionar as paixões (ou seja, objetivos e meta-objetivos) que deve servir, então a razão não é escrava dessas paixões, ou seja, a razão humana não é algorítmica.

Pensamos que o próprio Turing reconheceu que *esta lacuna da liberdade era essencial em sua concepção de uma máquina computacional*, mesmo se implementada por humanos realizando cálculos: "Um homem provido de papel, lápis e borracha, e sujeito à estrita disciplina, é de fato uma máquina universal" (TURING, 1948; *grifo nosso*). Notavelmente, aconteceu exatamente daquela maneira na organização interna dos grupos de trabalho de Bletchley Park criados para decodificar as mensagens criptografadas alemãs durante a Segunda Guerra Mundial (HINSLEY & STRIPP, 1993). Estar 'sujeito à estrita disciplina' significa não questionar de nenhuma forma as *regras e propósitos* do procedimento, isto é, da computação.

5. DETERMINAÇÃO, INDETERMINAÇÃO, AUTODETERMINAÇÃO

Mecanicismo em filosofia é a visão de que todos os seres, seja eles com vida ou sem vida, são como máquinas complicadas. Mecanicismo está intimamente conectado com o determinismo, desde que a revolução científica e tecnológica do século 17 fez alguns filósofos – Hume entre eles – acreditar que todos os fenômenos poderiam eventualmente ser explicados em termos de 'leis mecânicas', isto é, leis naturais governando o movimento e colisão de matéria sob a influência de forças físicas. Visões modernas mecanicistas de seres vivos, incluindo humanos, compreendem

processamento de informação mecânica como um elemento essencial da 'máquina viva', por exemplo nas teorias behavioristas de estímulo-resposta.

Computadores são um tipo especial de máquinas, chamados de máquinas algorítmicas, onde o aspecto do processamento de informação fica no centro; portanto nós queremos explorar o papel da determinação mecanicista em nosso modelo de inteligência, seja humana ou robótica.

Distinguimos três – ou talvez quatro – formas de relacionamento entre determinismo e o comportamento dos humanos e máquinas computacionais.

1. Hetero-determinação. O comportamento é integralmente determinado pelo estímulo recebido e pelo processamento computacional ou neurológico desse estímulo que passa a produzir uma resposta, de acordo com programas mais ou menos complexos e sistemas de avaliação. No caso dos computadores, o programador escreveu estes programas e sistemas de avaliação. No caso dos humanos, eles são baseados na genética, ou adquiridos a partir da educação e influência do ambiente. Este paradigma é uma evolução natural da concepção *Humeana* da natureza humana, onde humanos não são mais do que robôs biológicos complexos: *nós somos os escravos obedientes de nossas paixões, nossa biologia, nossa educação ou herança cultural*. Os paradoxos do comportamento determinístico têm sido ilustrados há muito tempo com o dilema do asno de Buridan, baseado nos escritos do filósofo francês Jean Buridan (c. 1295 – 1363), e mais recentemente nas histórias de ficção científica tais como *Runaround* (1942) por Isaac Asimov. O problema recebeu uma base matemática do cientista da computação Leslie Lamport, que o chamou de O princípio de Buridan (LAMPORT, 2012): "Uma decisão discreta baseada em uma entrada com um intervalo contínuo de valores não pode ser feita dentro de um período de tempo limitado". No artigo, originalmente escrito em 1984, no entanto não publicado até 2012, o autor também enfatiza as consequências de ignorar o princípio ao projetar dispositivos de engenharia.

2. Indeterminação. Essa visão complementa a anterior adicionando um certo grau de incerteza devido às causas físicas, seja pelo subsistema de avaliação (decisão por um fator de aleatoriedade) ou pelo subsistema de execução (o qual de fato significa que o sistema físico não se comporta exatamente como o comandado). A falta de determinação no

resultado do algoritmo não é uma surpresa para os cientistas da computação, os quais têm utilizado a aleatoriedade nos algoritmos por décadas (por exemplo algoritmos genéticos e outras técnicas de computação biologicamente inspiradas). Isso resolve o dilema do asno de Buridan pragmaticamente e faz com que o comportamento seja relativamente imprevisível (ainda que estatisticamente previsível). Todavia, o indeterminismo não adiciona nada essencialmente diferente à concepção *Humeana* da natureza humana. De fato, essas duas visões, hetero-determinismo e indeterminismo, concordam na negação radical da liberdade humana, que nós podemos observar na interpretação dos experimentos neurocientíficos realizados por Benjamin Libet e outros (LIBET et al., 1983; SOON et al., 2008): *a liberdade é uma ilusão*, afinal os atos voluntários são iniciados de maneira inconsciente no cérebro antes que o sujeito os perceba. Mesmo que tenham filósofos (incluindo Hume, no seu *Investigações de 1748*, ainda que não no seu Tratado de 1739) que considerem que a liberdade é compatível com o hetero-determinismo, pensamos que o senso comum de liberdade é precisamente o que a interpretação dos resultados experimentais expôs: se o comportamento é hetero-determinado ou indeterminado, a liberdade é uma ilusão. Autores contemporâneos como Daniel Dennett (DENNETT, 1991), também seguem uma abordagem compatibilista, mas na nossa visão os argumentos deles na verdade confirmam que eles pensam que a liberdade não é algo real que pode influenciar o comportamento humano, mas uma ilusão produzida pelo cérebro, seja ele determinístico ou indeterminístico.

3. Autodeterminação. Nesta posição as duas anteriores são rejeitadas. Se a liberdade humana, no senso comum, não é uma ilusão, então não é verdade que o comportamento humano é hetero-determinado (ainda que só estatisticamente) apenas pelo corpo material e o fenômeno que ocorre nele. Pelo contrário, ser realmente livre significa que os seres humanos se autodeterminam em suas ações. A autodeterminação admite duas versões:

> **a. Autodeterminação *direcionada aos* fins**. Nesta versão, os seres humanos são livres para buscar um certo objetivo final, e escolhem entre diferentes comportamentos visando alcançá-lo. Mas o objetivo final, como tal, já é dado. Isso implica em uma afirmação bastante modesta de liberdade, consistindo apenas na

(talvez computável) escolha entre vários meios de alcançar um fim dado, junto com uma opção mais substancial de seguir este fim ou não.

b. Autodeterminação *dos* fins. Nessa versão mais radical, os seres humanos não apenas se autodeterminam para um fim, mas autodeterminam os fins: os fins não são dados. É afirmado que humanos não apenas *tem* um destino (muito menos um destino trágico), mas eles próprios *forjam* seu destino. Um ser humano não apenas escolhe *como* vai se tornar algo, mas *o que* ele ou ela quer se tornar. E isso é precisamente o que torna difícil fazer certas escolhas (CHANG, 2013).

A autodeterminação possui dois difíceis problemas que não pretendemos resolver aqui. O primeiro, o problema metafísico mente-corpo, isto é, a interação entre o imaterial e o material (de qualquer forma, não consideramos que o Dualismo Cartesiano é uma solução válida). O segundo, o problema moral da arbitrariedade na autodeterminação dos fins: importa se alguém escolha este ou aquele fim para sua vida? Certos fins são reconhecidamente melhores que outros?

Deixando de lado estes dois problemas, alcançamos um ponto crítico para a visão *Humeana*-computacional dos seres humanos, tendo em vista que a autodeterminação não é uma função algoritmicamente programada (um algoritmo não pode autodeterminar seu propósito). Em outras palavras, comportamentos livres não são computáveis. Assim, se a autodeterminação é a verdadeira essência da liberdade humana, então seres humanos não são robôs *Humeanos*.

6. AS OBJEÇÕES DO PANPSICALISMO E DA PANLIBERDADE

Duas posições filosóficas foram revividas recentemente, que parecem poder refutar nosso argumento. O *panfisicalismo* diz que tudo no universo é consciente de uma maneira ou outra; em outras palavras, a consciência é uma característica fundamental no Universo, como David Chalmers coloca no *The Conscious Mind* (CHALMERS, 1996). *Panliberdade* é a versão do panfisicalismo, a qual declara que tudo no universo tem um certo grau de liberdade, até mesmo as partículas elementares da física. Esse é o resultado alegado pelo *Teorema do Livre-arbítrio* de Conway e Kochen (CONWAY & KOCHEN, 2006): dentro de certas suposições, se os humanos têm livre-arbítrio, no sentido de que o comportamento destes não é uma função computável do passado, então assim deve ser com as partículas elementares. Aparentemente, pode-se concluir isso, se partículas elementares são livres,

então é provável que os computadores também sejam; ou, se tudo é consciente, então computadores são conscientes, ainda que de uma forma diferente da dos seres humanos.

De fato, essas duas posições são controversas em si mesmas, com seus defensores e oponentes, como qualquer coisa que você possa imaginar que pode ser dito sobre liberdade e natureza humana (incluindo nossa própria posição). Visando a concisão do nosso argumento, não temos a intenção de refutá-las. Em vez disso, estamos satisfeitos com a demonstração de que, mesmo se o panfisicalismo ou a panliberdade forem propriedades reais do universo, nosso argumento permanece válido.

Ele é da seguinte forma. A essência de um computador algorítmico (lembre-se da definição de Turing do método computacional efetivo ou mecânico) é de dar conta da computação, isto é, obter o resultado desejado, alcançar o objetivo pré-definido. Se a computação falha em atingir seu objetivo, não é porque *é uma* computação, mas porque *não é*. Em outras palavras, é uma má implementação da computação ideal, ou é uma computação que não é capaz de superar a resistência do hardware de obedecer o software. Se o computador não computar como o esperado não pode ser atrelado ao indeterminismo desobediente das partículas elementares (seja pela sua indeterminação aleatória verdadeira, ou ao seu alegado livre-arbítrio incontrolável), assim não é um verdadeiro computador, ela é uma falha.

Assim, mesmo se nós assumirmos que as partículas elementares são livres, nós não poderíamos concluir que computadores são, pois o propósito geral de um computador é de neutralizar e controlar o livre (ou, pelo menos, indeterminístico) comportamento de seus componentes, de maneira a alcançar um objetivo. De fato, nós podemos argumentar de maneira similar quando os componentes são comumente reorganizados como seres livres. Lembre-se do computador feito por humanos que trabalhou no Bletchley Park, rodando programas projetados por Turing e outros, como citado acima: “Um homem provido de papel, lápis, e borracha, e sujeito à disciplina estrita, é de fato uma máquina universal” (TURING, 1948). As máquinas de Bletchley Park trabalharam como um computador apenas porque a iniciativa individual foi suprimida. Se um membro da equipe tivesse falhado em (ou não quisesse) cumprir sua subtarefa, então o Bletchley Park não mais teria operado como um algoritmo.

7. O TESTE DE TURING REVISITADO

Alan Turing propôs em 1950 seu famoso teste para determinar se uma máquina poderia ou não pensar, ou melhor, como um meio de definir como uma máquina pensante poderia ser (TURING, 1950):

> Eu proponho a consideração da seguinte questão, "Podem as máquinas pensar?" Isso deve começar com definições do que significa os termos 'máquina' e "pensamento'.

O teste é concebido como um procedimento metódico, um 'experimento' para dizer de uma maneira verificável se uma máquina pode ou não pensar. O restante do teste não requer que o papel de interrogador seja realizado por um humano, um grupo de humanos, ou mesmo uma máquina. De fato, a CAPTCHA (acrônimo para *Completely Automated Public Turing test to tell Computers and Humans Apart*) é um teste de Turing executado por uma máquina. Contudo, nós sabemos que não podemos distinguir algoritmicamente se uma sequência de eventos (isto é, o comportamento do sujeito sob exame) tem ou não algum propósito. Gregory Chaitin demonstrou (CHAITIN, 2005), uma derivação do Problema da Parada de Turing, que não há algoritmo que possa dizer inequívoca se uma sequência de números é determinística ou aleatória (isto é, com ou sem propósito).

Vamos supor um ser humano (assumindo que é verdadeiramente livre, isto é, autodeterminado) é sujeito a um Teste de Turing. O que aconteceria se o humano se propõe a mimetizar uma máquina e enganar o interrogador? Como pode o interrogador (humano ou mecânico, mas, de qualquer forma, metódico) defender a si mesmo do engano? Nós não sabemos o que um ser radicalmente livre sujeito ao teste fará. Tendo em vista que não é um objetivo pré-definido, ele pode decidir em falhar no teste. E se ele decide em seguir rigorosamente as instruções do interrogador, então não é um verdadeiro representante da humanidade desse momento em diante. Mesmo que não haja um desejo claro de enganar, se o imitador recebe instruções para "comportar-se como um humano", *o que significa, comportar-se como um humano?*

Por outro lado, uma máquina que é construída para passar no teste, isto é que mimetiza perfeitamente o comportamento humano (o jogo da imitação), tem o objetivo de imitar um ser… que não tem um objetivo *a priori*! Como pode ser assim? O que pode ser a especificação de fazer um comportamento não específico? Claro, nós podemos montar uma máquina que mimetiza o comportamento humano *típico*: de alguma forma errático, de alguma forma com propósitos, e assim por diante. Isso já foi alcançado em um certo grau com a tecnologia atual. Mas o ponto aqui é que *nós não podemos metodicamente (algoritmicamente), distinguir entre comportamentos com propósito e comportamentos sem propósito*, muito menos entre seres que se comportam de acordo com instruções dadas e seres que se comportam de acordo com instruções que eles próprios escolhem (isto é, de acordo com seus próprios propósitos). O comportamento típico (isto é, o comportamento médio), não é necessariamente o comportamento de um indivíduo. *Comportamento livre (autodeterminado) não é computável*.

Em outras palavras, podemos criar uma máquina (o imitador) que se comporta como humanos típicos, e podemos criar uma máquina (o testador) que distingue entre comportamentos humanos típicos e estranhos. Mas isso deixa de lado a questão essencial se este ser é ou não realmente livre, no sentido de que ele propõe seus próprios objetivos. Nós podemos construir robôs que desobedecem a seus donos humanos (BRIGGS & SCHEUTZ, 2015), mas ao fazer isso, estes robôs não fazem mais do que obedecer a seus programadores humanos. Criatividade, isto é, a capacidade de criar projetos e ir além de objetivos pré-determinados, não é uma propriedade verificável. Se seres humanos são, como Hume pressupunha, 'os escravos das paixões', então não há dificuldade a princípio em construir uma máquina imitadora, pois humanos são máquinas no fim das contas; e também não há dificuldade em criar uma máquina testadora, quando essas paixões são conhecidas (a especificação do teste pode ser 'humano' é aquele ser que busca essas paixões conhecidas). Mas se a autodeterminação é uma propriedade genuína dos humanos, então a estrada em direção à humanidade está fechada para máquinas computacionais. Neste sentido, nossa posição é uma objeção dupla (tanto da posição do imitador quanto do testador) para o Teste de Turing como um modelo de compreensão da inteligência humana, pois ele deixa de lado os aspectos não computacionais da liberdade e da escolha racional (claro, se a liberdade humana é negada, então nossa objeção é vazia). Nossa objeção, todavia, não diminui a importância do Teste de Turing como um programa de pesquisa para o desenvolvimento de Inteligência Artificial ou como critério para definir o que pode ser um pensamento artificial.

8. RESUMO DO ARGUMENTO

David Hume concebe a razão como escrava das paixões, o que implica que a razão humana possui objetivos predeterminados que não pode questionar. De outro lado, nós também estabelecemos que, por construção, um elemento essencial de um algoritmo rodando em uma máquina computacional (ou Máquina Lógica Computacional, como Alan Turing chamava) possui um propósito pré-determinado: um algoritmo não pode questionar seu propósito, pois se pudesse deixaria de ser um algoritmo.

Estabelecemos que H denota o conjunto de 'seres *Humeanos*', isto é, seres que seu comportamento obedece objetivos pré-determinados. E que A denota o conjunto de 'seres algorítmicos', isto é, seres cujo comportamento é algoritmicamente computado. Então A é o subconjunto de H, pois havendo objetivos pré-definidos é uma propriedade essencial da definição de A; portanto, ser A implica ser H, e não ser H implica em não ser A.

$$A \Rightarrow H$$

$$\neg H \Rightarrow \neg A$$

O inverso não é verdadeiro. Pode existir uma 'entidade dirigida por objetivos' (H) que o comportamento não é algorítmico ou mecânico (no sentido preciso de 'mecânico' de Turing), mas ainda ser fisicamente determinado (COPELAND, 2002).

Estabelecemos que S denota o conjunto de 'seres autodeterminados', isto é, seres que podem escolher livremente em seguir ou não um determinado fim, ou até podem autodeterminar os fins que querem seguir. Então, fica claro que ser S não implica em ser H, e dessa forma não ser A.

$$S \Rightarrow \neg H$$

$$\neg H \Rightarrow \neg A$$

$$\therefore S \Rightarrow \neg A$$

Por outro lado, ser A implica em não ser S.

$$A \Rightarrow H$$

$$H \Rightarrow \neg S$$

$$\therefore A \Rightarrow \neg S$$

Resumindo, se autodeterminação é essencial para a inteligência humana, então seres humanos não são nem seres *Humeanos*, nem máquinas algorítmicas.

Isso poderia acontecer, como Hume entendia, que humanos sempre agem de acordo com seus impulsos, desejos e vontades que não podem ser racionalmente desafiados. Isso pode ser facilmente imitado por máquinas. Também é concebível um sistema físico que de alguma forma é flexível o suficiente para autodeterminar seus próprios objetivos: nós humanos somos, sem dúvida, sistemas físicos, e nós somos capazes (ao menos na opinião de muitos, que não compartilham a visão *Humeana* da natureza humana) de autodeterminação. Não obstante, tais sistemas físicos, mesmo se forem produzidos por humanos, estão fora do paradigma computacional como foi definido classicamente por Turing e outros: *eles são sistemas físicos, mas não são, propriamente ditos, máquinas algorítmicas* que obedecem um programa (note que, como já havíamos explicado, a seleção de objetivo algorítmico continua sendo algorítmico, mesmo se for em um alto nível). Uma entidade capaz de autodeterminar seus próprios objetivos em alto nível, bem, não é algorítmica.

Neste ponto precisamos recapitular nossas suposições (isto é, essas afirmações que nós damos por garantidas sem precisas demonstrar ou, pelo menos, argumentar a favor delas), e mais especialmente nossas não-suposições, e distinguir ambas das nossas conclusões. Em particular: nós *não* assumimos que máquinas não podem ser livres; e, consequentemente, nós argumentamos que *se* humanos são livres, *então* humanos não são máquinas.

Por outro lado, o que realmente assumimos foram as definições de computador e liberdade. Nós definimos 'computador' como uma máquina algorítmica ou computacional (não é uma definição muito estranha, de fato, baseada na tradição da ciência da computação iniciada por Turing). Uma máquina algorítmica tem um objetivo pré-definido que foi desenvolvido para concluir; nós não inventamos essa propriedade. A definição de um computador como uma máquina algorítmica inclui, como um elemento essencial e princípio de desenvolvimento, o fato em que há um objetivo pré-definido. Desta forma (por força da definição dos termos) um computador não

pode mudar seu objetivo, porque assim não seria mais um computador, isto é, algo com um objetivo pré-definido. Isso seria simplesmente uma contradição: sem mais, nem menos.

Existem muitas maneiras de definir 'liberdade'. Definimos (também seguindo tradições bem estabelecidas) como autodeterminação, isto é, a possibilidade de escolher seus próprios objetivos, ou colocar hierarquia entre objetivos dados.

Certamente afirmamos que máquinas computacionais não possuem livre-arbítrio. Mas isso não é uma *suposição*: é uma *consequência* das definições de máquina computacional e liberdade, que é o ponto deste trabalho.

Todavia, notem que *não* afirmamos que é impossível 'fazer', 'construir', ou a forma que você quiser chamar isso, um ser livre artificial, seja orgânico, inorgânico (talvez eletrônico), ou misto. Apenas afirmamos que isso não será um robô, um computador ou uma máquina algorítmica. Isso não é apenas um problema com os computadores 'atuais'. Isso é um problema para os computadores (isto é, máquinas computacionais com objetivos pré-determinados) de todas as eras. Se um ser artificial vir a ser livre (isto é, autodeterminado), ele não será porque ele é um computador, mas porque ele é mais do que isso.

Se, em algum momento no futuro, conseguirmos fazer máquinas sem objetivos pré-definidos, nosso argumento não vai se aplicar, obviamente. De qualquer forma, nós ainda achamos que 'uma máquina sem um objetivo pré-definido' é um conceito contraditório: nós devemos inventar um novo nome para isso. Em um sentido, a reprodução humana já produz seres sem objetivos pré-definidos.

De alguma forma, nós admitimos que podemos construir seres artificiais capazes de alterar suas paixões, seus objetivos; por outro lado, nós não admitimos que esses seres artificiais podem ser propriamente chamados de 'máquinas computacionais'. Se a máquina de lavar pode decidir parar de lavar, então ela não é mais uma máquina de lavar. Se um robô pode selecionar seus próprios objetivos, então ele não é mais um robô, mesmo que ele continue sendo um dispositivo eletrônico. De fato, um dispositivo eletrônico, um ser físico, mas não um computador, não uma máquina computacional.

Esperamos que tenhamos clarificado a relação entre máquinas computacionais, razão humana e livre-arbítrio. Assumimos que a liberdade verdadeira, no seu sentido comum, requer autodeterminação, ao menos em uma forma fraca de 'autodeterminação *direcionada* para os fins' (mesmo se nós preferirmos a versão mais forte da autodeterminação *dos* fins).

De qualquer forma, *não* demonstramos que os seres humanos são realmente livres (autodeterminados). Apenas demonstramos que *se* humanos são livres, *então* eles não podem ser máquinas algorítmicas. Desta forma a inteligência humana, assim entendida, não pode ser apropriadamente definida como processos algorítmicos, e o comportamento humano não pode ser emulado por robôs algorítmicos. Se nós, em algum futuro incerto, pudermos produzir em nossos laboratórios um tipo de robôs não algorítmicos que podem ser apropriadamente chamados de livres, e se ainda eles poderão ser chamados de robôs, será objeto de mais pesquisas.

REFERÊNCIAS

BRIGGS, G., Scheutz, M. **"Sorry, I can't do that:"** Developing mechanisms to appropriately reject directives in human-robot interactions. Proceedings of the 2015 AAAI Fall Symposium on AI and HRI. Available at http://www.aaai.org/ocs/index.php/FSS/FSS15/paper/download/11709/11522.

CHAITIN, G. **Meta Math! The Quest for Omega.** New York: Vintage Books, 2005.

CHALMERS, D. **The Conscious Mind: In Search of a Fundamental Theory.** Oxford: Oxford University Press, 1996.

CHANG, R. Grounding practical normativity: going hybrid. **Philosophical Studies** 164(1): 163-187. 2013. DOI: 10.1007/s11098-013-0092-z

CONWAY, J., KOCHEN, S. The Free Will Theorem. **Foundations of Physics** 36(10): 1441–1473, 2006. DOI: 10.1007/s10701-006-9068-6.

COPELAND, B.J. The Church-Turing Thesis. **The Stanford Encyclopedia of Philosophy** (Summer 2015 Edition), Edward N. Zalta (ed.). 2002. The text is available online at http://plato.stanford.edu/archives/sum2015/entries/church-turing.

COPLESTON, F. **A History of Philosophy**, vol. 6, pp. 405–406. London: A&C Black. 1999.

DENNETT, D.C. **Consciousness Explained**. London: Penguin Books, 1991.

HILL, R.K. What an algorithm is. **Philosophy & Technology** 29(1): 35–59, 2015.

HINSLEY, F. H., STRIPP, A., eds. **Codebreakers: The inside story of Bletchley Park.** Oxford: Oxford University Press, 1993.

HUME, D. **A Treatise of Human Nature: Being an Attempt to introduce the experimental Method of Reasoning into Moral Subjects**, II-iii-3, London: John Noon, 1739. It was later reworked and

published in 1748 as An Enquiry Concerning Human Understanding. 1739. The text is available online at http://www.gutenberg.org/ebooks/4705 and http://www.davidhume.org/texts/thn.html.

KNUTH, D.E. **The art of computer programming**, volume 1 (3rd Ed.): fundamental algorithms. Redwood City: Addison Wesley Longman Publishing Co. 1997.

KROES, P. Engineering and the dual nature of technical artefacts. **Cambridge Journal of Economics** 34(1): 51-62, 2010. DOI: 10.1093/cje/bep019.

LAMPORT, L. Buridan's Principle. **Foundations of Physics** 42(8): 1056-1066, 2012. DOI: 10.1007/s10701-012-9647-7.

LIBET, B., GLEASON, C.A., WRIGHT, E.W., PEARL, D.K. Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential) - The Unconscious Initiation of a Freely Voluntary Act. **Brain** 106(3): 623–642. 1983. DOI: 10.1093/brain/106.3.623.

SOON, C.S., BRASS, M., HEINZE, H.J., HAYNES, J.D. Unconscious determinants of free decisions in the human brain. **Nature Neuroscience** 11(5): 543–545. 2008. DOI: 10.1038/nn.2112.

TURING, A.M. On computable numbers, with an application to the Entscheidungsproblem, **Proceedings of the London Mathematical Society** 2(42): 230–265. 1936.

TURING, A.M. Intelligent Machinery. National Physical Laboratory Report. In Meltzer, B., Michie, D. (eds), **Machine Intelligence** 5. Edinburgh: Edinburgh University Press, 1969. Digital facsimile viewable at http://www.AlanTuring.net/intelligent_machinery. (1948).

TURING, A.M. Computing machinery and intelligence. **Mind** 59: 433-460. 1950.

VARDI, M. What is an algorithm? **Communications of the ACM** 55(3): 5–5, 2012. DOI: 10.1145/2093548.2093549.

FAGUNDES, Ismail. BERNARDI, Mariana. SÃO OS SERES HUMANOS ROBÔS HUMEANOS?. *Kalagatos*, Fortaleza, vol.19, n.2, 2022, eK22028, p. 01-19.

Recebido: 05/2022
Aprovado: 06/2022