POSSIBILIDADES DE AUTONOMIA DA INTELIGÊNCIA ARTIFICIAL

André Guimarães[1]

https://orcid.org/0000-0003-1856-9839

https://doi.org/10.33871/27639657.2024.4.1.9192

RESUMO: O presente artigo busca refletir sobre possibilidades de autonomia da IA, a partir de argumentos favoráveis e argumentos contrários, buscando manter um equilíbrio entre o tecnoentusiasmo e o tecnocatastrofismo. o artigo deixa registrada uma vereda sobre possibilidades de singularização das máquinas digitais, a partir da discussão dos estados objetais não objetiváveis. A instanciação material da Máquina de Turing (conceito abstrato) traria a possibilidade, teórica, de construtos identitários singulares a máquinas digitais. A existência dessa possibilidade deixa aberto o caminho para que as máquinas digitais se tornem autônomas e inteligentes.
Palavras-chave: Filosofia da Mente. Inteligência Artificial. Autonomia da Inteligência Artificial.

POSSIBILITIES OF AUTONOMY OF ARTIFICIAL INTELLIGENCE

ABSTRACT: The paper reflects on possibilities for AI autonomy, based on favorable and unfavorable arguments, seekint to keep a balance betwenn technoenthusiasm and technocatastrophism. The paper provides a perspective on the possibilities of singularizing digital machines, based on the discussion of non-objectivable object states. The material instantiation of the Turing Machine (abstract concept) would bring the theoretical possibility of singular identity constructions for digital machines. The existence of this possibility leaves the way open for digital machines to become autonomous and intelligent.
Keywords: Mind Philosophy. Artificial Intelligence. Artificial Intelligence Autonomy.

INTRODUÇÃO

Os avanços recentes da Inteligência Artificial (IA) despertaram manifestações de preocupação com os possíveis desdobramentos dessa evolução para o futuro da humanidade. O presente artigo busca refletir sobre possibilidades de autonomia da IA, a partir

[1] Doutor em Filosofia pela UFSCar, Mestre em Comunicação pela UMESP, Mestre em Informática pela PUC-Campinas, graduado em Economia pela UFMG. Docente do Mestrado em Poder Legislativo do Centro de Formação, Treinamento e Aperfeiçoamento da Câmara dos Deputados e docente do Mestrado em Gestão Estrat´égica de Organizações, do IESB.

de argumentos favoráveis e argumentos contrários, buscando manter um equilíbrio entre o tecnoentusiasmo e o tecnocatastrofismo.

. Compreender a IA como realmente autônoma implica assumir a possibilidade de um novo tipo de artefato, que na realidade deixaria de ser artefato para se constituir em nova entidade, um *Outro*, o qual o homem precisará entender e estabelecer um *modus vivendi*. Essa convivência, por sua vez, acontecerá em um novo espaço, alterado pela presença dessa nova entidade.

O artefato intangível e plenamente inserido na subjetividade do humano, extrai desse qualidades intrínsecas e adquire pretensões de autonomia. O artigo, portanto, discute um dos temas mais polêmicos da relação entre homens e máquinas. Confortável enquanto relação de servidão (a máquina servindo o homem), essa relação começa a incomodar quando se pretende isonômica (máquina e homem colaborando) e torna-se temível com a perspectiva de inversão dos papeis (o homem servindo à máquina).

A INTELIGÊNCIA SEM FRONTEIRAS

A pretensão de simular a inteligência humana é antiga e suas origens não são possíveis de serem rastreadas na história. Há indícios de preocupação com autômatos já no pensamento grego. Porém, a IA como um campo de pesquisa delineou-se a partir do final da década de 1950, quando também surgiam e estavam em franco desenvolvimento os primeiros computadores digitais. Por suas peculiaridades, os computadores digitais, como máquinas de finalidades gerais, baseados na Máquina de Turing, representavam, pela primeira vez, a possibilidade real de materialização da inteligência humana em outro tipo de mídia, que não o tecido cerebral humano.

Em seu início, a IA mesclava a abordagem da então incipiente ciência cognitiva com a ciência da computação e tinha como propósito a criação de modelos computacionais para a compreensão da cognição humana. Nas duas décadas iniciais de seu desenvolvimento, a Inteligência Artificial assumiu como projeto a construção de softwares que teriam a

capacidade de igualar o comportamento humano inteligente. Posteriormente, essa linha inicial de pesquisa veio a ser chamada de GOFAI – *Good Old-fashioned Artificial Intelligence*[2].

Em seu fundamento filosófico, a GOFAI assumiu um controverso posicionamento entre o cartesianismo dualista e o materialismo monista, que Floridi (1999) chamou de materialismo computacional. Para Floridi, essa posição estabelece que a inteligência é "biologicamente independente do corpo e a-social, mas também completamente independente da mente e, portanto, implementável por um (sem cérebro, sem mente e sem vida) sistema lógico-simbólico de finalidades gerais"(Floridi, 1999, p. 133).

Considerada como algo independente do corpo e essencialmente individual, a concepção de inteligência mantém uma perspectiva dualista, fortemente criticada pelas posições mais recentes das ciências cognitivas. Como independente da mente e passível de ser implementada em outros dispositivos, que alcancem os mesmos resultados, por meio de processos inteiramente diversos, a inteligência é compreendida de forma materialista. A combinação das duas perspectivas, originariamente almejada pelo chamado materialismo computacional, revela-se uma impossibilidade teórica e prática.

Sustentar o materialismo computacional, portanto, significava aceitar uma vertente funcionalista combinada a um reducionismo, que iguala a inteligência à computabilidade. Essa redução se torna possível mediante a igualdade primeira entre inteligência e raciocínio, e entre o raciocínio e o processamento de símbolos, em um segundo momento. Uma das grandes dificuldades dos pesquisadores da GOFAI foi deixar de entender que [inteligência = raciocínio = processamento de símbolos = computação] era um reducionismo e não uma equação a ser entendida literalmente. Teixeira corrobora essa visão, ao argumentar que "por trás da GOFAI está o paradigma simbólico, ou seja, a noção de que a mente é um sistema formal que manipula símbolos (representações) através de programas computacionais que resolvem problemas" (TEIXEIRA, 2005, p. 35).

Apesar dessas dificuldades conceituais, a GOFAI foi aplicada com êxito em diversas áreas, como demonstração e prova de teoremas, jogos, planejamento comportamental de robôs por meio de análises de meios e fins, sistemas especialistas, percepção acústica e visual

[2] A sigla foi criada em 1981 pelo filósofo J. Haugeland e significa, em uma tradução literal, a *boa e velha inteligência artificial*.

e reconhecimento de padrões. Todas essas áreas apresentam alguns pontos em comum: são computáveis, independentes em relação à experiência, ao corpo e ao contexto. Esses pontos em comum não são devidos ao acaso, mas decorrem do fato de que um computador é capaz de realizar tarefas inteligentes desde que seja capaz de internalizar todos os dados relevantes. Por essa razão, as aplicações da GOFAI são limitadas à domínios muito restritos, a partir dos quais os programadores criam micromundos. Esses, por sua vez, representam uma combinação "dos compromissos ontológicos que os programadores assumem quando concebem o sistema e que desejam que o sistema adote" (FLORIDI, 1999, p. 146).

Essa forte restrição de domínio leva a GOFAI a um paradoxo: quanto mais restrito o domínio e, portanto, mais passível de formalização, mais viável é o desenvolvimento de aplicações, porém menos inteligentes parecerão as mesmas[3]. Ou seja, na verdade não há uma *inteligentificação* das máquinas, mas sim uma *estupidificação* da inteligência. Uma não-restrição do domínio, contudo, leva a problemas insuperáveis para a IA, tais como a explosão combinatorial e a rigidez de estrutura.

Com o tempo, surgiu uma nova abordagem no campo de pesquisa, que veio a ser conhecida como *Light Artificial Intelligence* – LAI. Ao invés de se propor a construir hardwares e softwares para igualar a inteligência, a LAI busca se orientar para a consecução das tarefas e a resolução dos problemas. Assim, a pesquisa em IA tenta se desvencilhar dos resquícios do dualismo cartesiano, por meio de uma abordagem mais estritamente funcionalista, a qual abrange a compreensão de que diferentes tarefas podem ser realizadas de modos muito distintos. No nascedouro da LAI estava a concepção de que "tarefas inteligentes poderiam ser realizadas por dispositivos que não teriam a mesma arquitetura nem a mesma composição biológica e físico-química do cérebro humano" (TEIXEIRA, 2004, p. 60). A questão essencial, então, passa a ser se existe uma forma computacional de resolver uma determinada tarefa. Ao invés de *estupidificar* a inteligência, trata-se de *estupidificar* o processo por meio do qual se resolve o problema.

[3] Uma das grandes queixas dos defensores da IA é justamente que a cada nova conquista da mesma, seus antagonistas reagem dizendo que o que se conseguiu, na verdade, não tem a ver com inteligência.

A nova abordagem representou um grande avanço, permitindo o desenvolvimento de aplicações de IA ainda mais bem-sucedidas e utilizadas em uma variedade maior de problemas. No entanto, para alcançarem sucesso, as aplicações de IA continuaram restritas a lidar com problemas claramente definidos, tarefas que sejam redutíveis a seqüências de procedimentos heurísticos com propósitos específicos e instruções repetitivas. Na visão de Floridi (1999), isso se deve à própria natureza dos computadores, que operam basicamente por meio de sua capacidade de detectar e processar uma relação diferencial, usualmente binária, e proceder inferencialmente a partir dessa base. Segundo esse autor, "nós precisamos não esquecer que apenas sob condições especialmente determinadas uma coleção de relações diferenciais detectadas, concernentes a algum aspecto empírico da realidade, pode substituir o conhecimento experiencial direto desse" (FLORIDI, 1999, P. 215).

Nem toda situação experiencial – nem todo o conhecimento gerado pelas mesmas – é passível de ser traduzida em relações diferenciais binárias, que são ordinariamente empregadas pelos computadores digitais em seus processos inferenciais. A questão essencial e relevante, conforme Churchland, torna-se, então: "se as atividades que constituem a inteligência consciente são, todas elas, algum tipo de procedimento computacional" (CHURCHLAND, 2004, p. 171).

Existe um limite na computabilidade, relacionado diretamente à possibilidade de desenvolvimento de um algoritmo para a resolução dos problemas, uma vez que há problemas que não podem ser homogeneizados por estados definidos e, por conseguinte, não são tratáveis algoritmicamente. O próprio conceito de inteligência permanece, assumidamente, vago, o que torna ainda mais difícil a tarefa de reconhecer uma inteligência artificial. Porém, caso se adote uma linha funcionalista, de buscar as três características mais comumente associadas à inteligência (compreensão, capacidade de solução de problemas e aprendizagem), qualquer entidade que apresente esses três atributos pode reivindicar o status de inteligente. Essa é a vereda que vem sendo percorrida pelos pesquisadores de IA. Ao invés de se engalfinharem na controvérsia sobre a natureza da inteligência humana, definem suas metas em termos empíricos e operacionais. Nessa perspectiva, funcionalista, uma definição de inteligência é desnecessária.

IA E A NOÇÃO DE AGENTE

No sentido aristotélico, agente é o que faz a ação. Interessante observar que Aristóteles, em sua Poética, admitia a possibilidade de uma peça sem personagens, mas não a de uma peça sem ação, o que significa que o papel de agente não precisaria necessariamente ser incorporado por um personagem. Aristóteles delineou quatro critérios para personagens dramáticas, que podem ser aplicados a softwares[4]. O primeiro critério é ser *bom*, ou virtuoso, no sentido de completar com êxito sua função. Bons personagens fazem o que seus criadores desejam que eles façam, no contexto do drama que se desenrola. O segundo critério é ser *apropriado* às ações que realiza. O terceiro critério é que os agentes sejam *como reais*, no sentido de que existam conexões causais entre seus traços e suas ações. O quarto critério é que o agente seja *consistente*, ou seja, não apresente mudanças arbitrárias em seu comportamento.

Adotando-se essa perspectiva, qualquer programa de computador que execute uma ação pode ser considerado um agente. No cotidiano, em sua relação com os computadores, as pessoas já se acostumaram a lhes atribuir pressupostos de agência – "eu fiz isso e o computador travou" ou "o processador de textos perdeu toda a formatação do documento". Esse comportamento pode ser explicado nos moldes colocados por Dennett, quando esse afirma que em relação a máquinas complexas a melhor postura a ser adotada é a postura intencional.

Em termos sociais e legais, um agente é algo ou alguém que tem o poder de agir no lugar de outrem. Essa perspectiva levanta a questão da responsabilidade. E já começa a trazer discussões inusitadas. A seguir, apresentamos argumentos favoráveis e contrários à possibilidade de que a IA passe a ser considerada plenamente como agente.

ARGUMENTOS CONTRÁRIOS

[4] Essa possibilidade foi levantada por LAUREL (1993).

Dennett apresenta três posturas possíveis perante o mundo e as coisas. A postura objetal, a postura de design e a postura intencional. Para esse autor, deve-se tratar como agente "de fato um agente racional, cujas crenças básicas, desejos e outros estados mentais que exibem intencionalidade ou '*aboutness'* e cujas ações possam ser explicadas (ou previstas) com base no conteúdo desses estados" (DENNETT, 1991, p. 76, tradução livre). A IA, por sua natureza, candidata-se quase que naturalmente a ser tratada pela postura intencional.

Boden (1981, p. 145) argumenta que a postura intencional decorre de uma analogia profunda entre a forma de funcionamento do organismo humano e a da máquina. As intenções, no orgânico, teriam a função de controlar as operações corporais executadas para atender aos propósitos valorizados pelo agente. Uma teoria de cunho psicológico para entender a intencionalidade precisaria iniciar pela especificação das micro-operações corporais básicas a partir das quais surgem os macro-efeitos intencionais. As unidades corporais responsáveis pela execução dessas micro-ações são simples e procedurais, e as realizam de forma automática. Tanto o orgânico quanto o inorgânico seriam capazes de ser palco para a emergência de comportamento complexo a partir da interação de uma multiplicidade de micro-agentes, seguindo regras muito simples.

Na linha do pensamento de Boden,

> por analogia, uma instrução de computador de alto nível (em uma linguagem de programação) pode ser analisada em uma série de instruções em código de máquina, mas se alguém perguntar como qualquer uma *dessas* é efetivada, a única resposta possível se dará em termos eletrônicos (e não programáticos) (BODEN, 1981, p. 146, tradução livre).

Há quem enxergue nos processos emergentes uma via para o surgimento da semântica a partir da sintática, uma tendência de inspiração antiga. Descartes, analisando a formação de sentido das palavras, percorre um caminho que vai de mecanismos puramente sintáticos (excitações da glândula pineal) ao semântico:

> as palavras, que excitam na glândula movimentos, os quais segundo a instituição da natureza, representam para a alma somente o som delas, quando são proferidas com a voz, ou com a figura de suas letras, quando são escritas, e que no entanto, pelo hábito que adquirimos pensando no que elas significam quando ouvimos seu som ou então quando vemos suas letras, costumam levar a conceber esse

> significado e não a figura de suas letras ou o som de suas sílabas (DESCARTES, 1998, p. 65).

Os programas podem servir para explicar como as intenções surgem, com como a forma pela qual os efeitos complexos da intencionalidade se compõem a partir das operações de nível mais baixo, mas somente o hardware pode explicar satisfatoriamente a base causal das intenções.

A questão da intencionalidade está no centro da discussão quanto à adequação de se atribuir ou não status de agente aos computadores e programas. Os argumentos dos que são contrários a essa *heresia* fundamentam-se no fato de que os computadores são construídos deliberadamente para funcionar como incorporações de programas (sistemas representacionais) cujo sentido é atribuído pelos seres humanos. Portanto, qualquer eventual *interesse* do computador não seria intrínseco à sua natureza, mas sempre um interesse *parasitário* do interesse humano. Apesar de apresentar comportamento inteligente, como não tem uma inteligência genuína subjacente, o computador seria um produtor de efeitos sem causas.

Muitos dos defensores da IA escolhem ignorar a questão da (falta de) intencionalidade, que remete à existência de um sujeito consciente, vivo, que pensa, calcula, escolhe, age e persegue objetivos porque tem necessidades, desejos, temores, esperanças, prazeres. Na base do humano há um sentimento profundamente enraizado, de falta, "sentimento de incompletude, está sempre a vir para ele, incapaz de coincidir com o si na plenitude imóvel do ser que é o que é" (GORZ, 2005, p. 92).

O computador seria um dispositivo exclusivamente procedural, que executa suas tarefas apenas quando disparado pelo usuário, cujo comando gera uma série estável de alterações em impulsos elétricos, que são então traduzidos pelo software. Turing já sinalizava para o problema:

> quando um computador humano está trabalhando em um problema ele pode usualmente aplicar uma dose de senso comum para ter uma ideia de quão apuradas são suas respostas. Com um computador digital, nós não podemos mais contar com o senso comum, e os limites dos erros precisam ser baseados em algumas desigualdades provadas (TURING, 2004,p. 391, tradução livre).

A abordagem computacional tenderia a se apegar ao formalismo, às representações simbólicas e às referências lógicas, buscando a certeza, a correção, a completude e o detalhe, ao mesmo tempo em que elimina a ambiguidade. Por causa desse alto grau de formalização e abstração, "o âmbito de intermediação entre ideia e resultado é completamente compreendido no interior da dimensão simbólico-racional, na qual devem ser utilizadas uma operatividade lógico-matemática" (CAPUCCI, 1997, p. 131).

O computador executa operações sobre sinais sem evocar as idéias correspondentes, uma espécie de pensamento cego. Nos dizeres de Gorz,

> trata-se de um 'pensar sem pensamento', ou seja, de um pensamento que não precisa ser penado nem entendido por nenhum sujeito, pois funciona como uma 'máquina simbólica', cujos modos de operação simbolicamente cifráveis, realmente, provocam, sem rodeios por consciências, efeitos diretos no real (GORZ, 2005, p. 83).

No suposto pensamento da máquina estariam ausentes o sujeito, a percepção, a referência a objetos exteriores passíveis de representação ou presentificação. É um pensamento livre das amarras – internas e externas – da experiência, operando apenas com signos e suas relações. A máquina computacional, operando às cegas, é incapaz de recuar para fora da tarefa em execução e examinar o que já foi feito, restando impossibilitada de notar mesmo os fatos mais óbvios a respeito do que está fazendo. Segundo Hofstadter, "a diferença, portanto, é a de que é possível para uma máquina agir sem observar; e é impossível para um ser humano agir sem observar" (HOFSTADTER, 2001, p. 42).

Quando o autômato cego executa o algoritmo, a potência operatória passa ao primeiro plano. Norman defende que os processos de pensamento dos humanos não são como a lógica matemática das máquinas: "na verdade, se os processos de pensamento dos humanos fossem como os da lógica, nós não teríamos precisado inventar a lógica como uma ajuda ao pensamento. A lógica é importante *porque* ela é diferente" (NORMAN, 1993, p. 228, tradução livre).

Os processos da lógica formal ignoram conteúdo e contexto (pensamento cego que opera sobre representações simbólicas), enquanto que o pensamento humano trabalha juntamente o contexto e o conteúdo dos problemas. De fato, a lógica, em uma acepção

técnica, não se refere à racionalidade em geral, mas à inferência da verdade de uma afirmação a partir da verdade de outras afirmações com base apenas na forma destas e não no conteúdo. Leibniz, no *Accessio ad arithmeticam infinitorum*, aplica uma situação análoga em pessoas, quando exemplifica que quando alguém diz um milhão, não consegue imaginar todas as unidades daquele número, porém é capaz de fazer cálculos exatos com base nessa cifra.

Essa perspectiva encontra-se também na objeção de Heidegger à proposta da lógica de inspiração booleana de conectar proposições ignorando sua dimensão semântica. Os métodos matemáticos, para Heidegger, permitiram a construção de um sistema de ligação de enunciados, razão pela qual se denominou essa lógica de *lógica matemática*. Heidegger afirma que os propósitos da lógica matemática são possíveis e legítimos, porém essa deve ser entendida como "uma coisa de completamente diferente de uma lógica, quer dizer, de uma reflexão sobre o λόγος [logos, em grego no original]" (HEIDEGGER, 1987, p. 154). Ainda segundo o pensamento de Heidegger,

> a própria logística é antes e somente uma matemática aplicada a proposições e a formas de proposição. Toda a lógica matemática e a logística se colocam necessariamente no exterior desse domínio da lógica porque, de acordo com os seus próprios fins, a logística deve utilizar o λόγος, o enunciado, como mera ligação de representações, quer dizer, de uma forma fundamentalmente insuficiente (HEIDEGGER, 1987, p. 154).

A única centelha de inteligência que é atribuída de forma unânime ao computador é sua capacidade de discriminar entre diferenças binárias e ser capaz de se comportar logicamente com base nessa distinção. No nível mais básico, o sistema é físico, sem qualquer tipo de representação explícita, apenas fenômenos físicos. Esses vão adquirir um significado apenas no nível mais alto do sistema lógico, no qual se encontram, por exemplo, as portas OR, ou a interpretação de uma presença/ausência de voltagem como 1/0. O sistema lógico é uma primeira abstração derivada do sistema físico e tornará possível, em outro nível, o sistema conceitual, representado pelas aplicações de software e linguagens de programação.

As descrições de um mesmo processo, em níveis diferentes, são muito distintas entre si e apenas os níveis mais elevados encontram-se aptos a serem compreensíveis por humanos. Lévy afirma que

> é preciso insistir no fato de que os processamentos em questão são sempre operações físicas elementares sobre os representantes físicos dos 0 e 1; apagamento, substituição, separação, ordenação, desvio para determinado endereço de gravação ou canal de transmissão (LÉVY, 1999, P. 151).

A natureza dos processos computacionais é sintática, o que faz com que os dispositivos que disponibilizam informação para esses processos sejam responsáveis tanto por seu formato quanto por sua qualidade. Essa natureza sintática está presente também nos mecanismos de armazenagem dos computadores. A máquina acumula registros de bytes, copiados com fidelidade total. Um processo inteiramente formalizável e reprodutível ao infinito. Mas uma ínfima discrepância na cópia digital pode inviabilizar sua reprodução. A memória humana é mais voltada para manter as relações importantes no mundo (padrões), independentemente dos detalhes. O sistema de memória do humano armazena sequências de padrões, os quais são recuperados de modo auto-associativo. Os padrões são armazenados no cérebro em formato invariante e em uma hierarquia. O modo auto-associativo de recuperação está na base da competência do sistema nervoso central para recuperar padrões completos, mesmo quando diante de dados sensoriais parciais ou distorcidos. Um computador não armazena, normalmente, sequências de padrões, embora alguns recursos de softwares atuais permitam uma simulação desse comportamento. Mesmo assim, todavia, memórias auto associativas artificiais falham em reconhecer padrões caso eles sejam movidos, rotacionados, sofram mudanças de escala ou qualquer outra transformação.

Ainda que a máquina computacional tenha capacidades autopoiéticas, funcione com processamento em paralelo, de modo randômico e sem distinção exata entre hardware e software, os processos por ela executados são cálculos. Cálculo implica manipulação e recombinação de símbolos atômicos, por meio de operações discretas e descontínuas, sem que haja a possibilidade de se determinar um estado intermediário entre o estado atual e o imediatamente posterior. Dado ainda que o alfabeto de símbolos atômicos sobre os quais são executadas operações é necessariamente finito, a máquina digital é determinista por construção. Os computadores operam procedimentos efetivos, os algoritmos. Maner (2002) levanta que um algoritmo vai infalivelmente gerar o resultado desejado após um número finito de passos, se receber entradas válidas suficientes. Nisso, o procedimento algorítmico é

diferente do heurístico, que opera pulando procedimentos, que tendem a produzir o resultado desejado quando obtém a entrada certa.

Lévy (1998) chama a atenção para a necessidade de se distinguir entre determinismo e previsibilidade. O formalismo algorítmico define implicitamente suas relações computacionais por meio da totalidade de suas relações computacionais com todos os outros estados do sistema em questão (por exemplo, relações de sucessão). Determinismo refere-se ao postulado de que dado o estado de um sistema em um determinado instante, o estado desse sistema em todo momento ulterior é determinado pelo movimento de suas partes. Previsibilidade refere-se à possibilidade de se prever efetivamente qual será a evolução de um sistema qualquer. Lévy argumenta, contudo, que isso não significa que processos biológicos, de padrão contínuo, não possam ser simulados por algoritmos apropriados, embora deixe uma ressalva: "mas admitir a possibilidade de representar um processo através de um cálculo é uma coisa; pretender que *é* um cálculo é outra" (LÉVY, 1998, p. 126).

Argumenta-se, ainda, a constrição das máquinas computacionais aos compromissos ontológicos assumidos pelos programadores, tanto aqueles implícitos que assumem quando escrevendo um programa, quanto aqueles que desejam permitir que o sistema adote livremente. Essa restrição das máquinas computacionais aos micro-mundos criados por seus programadores seria uma das causas da limitação na quantidade de respostas possíveis a serem dadas pelos sistemas, diante de variações do ambiente. A inteligência humana, por sua vez, é notória por sua capacidade de equilibrar respostas criativas a mudanças no ambiente com a possibilidade de se desligar do mesmo (transcendência).

Beavers entende que essa restrição das máquinas computacionais tem raízes mais profundas, brotando a partir da própria limitação da lógica binária. Para esse autor, apenas em condições muito especiais a detecção de relações diferenciais (binárias, por exemplo) com relação a algum aspecto empírico da realidade pode substituir o conhecimento experencial e direto sobre o mesmo. Por isso, "os computadores podem ser infalíveis para ler um código de barras, mas não podem explicar a diferença entre uma pintura de Monet e uma de Pissarro" (BEAVERS, 2002, p. 69, tradução livre). Há um diálogo quase lendário, atribuído a Picasso – quando questionado por oficiais franquistas sobre a obra Guernica – "Você fez isso?", teria respondido – "Não, vocês fizeram". Um computador jamais seria capaz de entender o sentido

desse diálogo, porque há uma natureza cumulativa e irreversível no conhecimento, na experiência e no engajamento corporal.

Pollock (2000) desenvolveu uma crítica a partir das características inerentes à inteligência humana. Para esse autor, nossa inteligência é sincronicamente defensável, pois uma proposição pode ser garantida em relação a um conjunto de entradas perceptuais e não garantida em relação a um conjunto mais amplo de entradas. E é diacronicamente defensável, pois uma proposição pode ser justificada em um estágio de raciocínio e injustificada em outro estágio posterior, sem o acréscimo de nenhuma entrada perceptual. O que Pollock tentou demonstrar foi a capacidade humana de lidar com paradoxos e contradições, que não seria aplicável a sistemas exclusivamente baseados em lógica formal.

A adaptabilidade é uma característica fundamental dos organismos vivos. A capacidade de responder apropriadamente, em uma variedade indefinida de formas, à imprevisível[5] variedade de contingências. Ser capaz de lidar com as contingências envolve procedimentos pelos quais uma situação nova é mapeada em uma estrutura representacional pré-existente, causando mudanças na mesma. Essas mudanças comportamentais ocorrem não somente em função de mudanças no ambiente, mas também em decorrência da compreensão daquilo que outros esperam de nós, a interpretação da intencionalidade de terceiros, ou uma intencionalidade de segundo grau (em relação ao indivíduo). Dreyfus usa o exemplo da linguagem para embasar esse tipo de argumentação contra os agentes informatizados – "aprender uma linguagem não é apenas aprender um conjunto fixo de palavras e construções gramaticais, mas usar esse equipamento lingüístico em situações sempre novas" (DREYfUS, 2000, p. 203, tradução livre). A linguagem, nessa acepção, é uma computação ao infinito. O que parece ser uma tradução possível da argumentação de Dreyfus é que se não houvesse limite de tempo no teste de Turing [tempo = ∞], os computadores nunca passariam no teste.

Um agente com capacidades adaptativas precisa ser um agente autodirigido. Embora existam regras que governam os processos de transformação dos dados sensoriais em estados conscientes – as quais ensejariam descrições e reproduções algorítmicas – uma das tarefas

[5] Pelo menos do ponto de vista do organismo.

mais difíceis da robótica atual é especificar uma tarefa para execução diante da imprevisibilidade do ambiente. Na raiz desse problema estaria uma diferença na orientação primária de organismos biológicos, os quais, ao invés de serem orientados para a tarefa (*task specified*), são orientados para o comportamento (*behavior specified*). Uma orientação para a tarefa requer uma especificação procedural rígida (o *o que* e o *como*), enquanto que na orientação para comportamento sabe-se *o que,* mas resolve-se o *como* em tempo real, no momento em que o organismo interage com o ambiente. Fundamental para essa performance em tempo real é o sentido da propriocepção.

Há os que argumentam que o caráter funcionalista da teoria computacional da mente lhe daria uma condição de meramente substitutiva, não explicativa. As máquinas computacionais são construídas não para explicar o pensamento, mas para substituí-lo, quando o esforço de o empregar é penoso. Conforme Pinto, o extraordinário valor prático dos computadores decorre justamente de sua absoluta inutilidade teórica: "se um computador imita algum comportamento racional inteligente de um homem, no máximo, tem o valor da substituição de um segundo homem no primeiro" (PINTO, 2005, p. 23).

Dreyfus também contesta esse suposto isodinamismo entre humanos e máquinas computacionais:

> não tenho nada contra a idéia de que o computador possa ser inteligente, contesto somente a hipótese dos 'sistemas de símbolos físicos', ou seja, a teoria segundo a qual nós, humanos e computadores, somos duas 'espécies' da mesma 'raça', em especial daquela que utiliza 'símbolos' para representar o mundo exterior (DREYFUS, 1993, p. 210).

Igualdade de comportamento não é igualdade de essência. Pinto corrobora a negação do isodinamismo, com termos fortes: "a ironia, noção filosófica, converte-se em estelionato, figura jurídica, quando se pretende impingir por equivalente o simulacro artificial de um ato biológico executado pela matéria viva por força de uma necessidade imperiosa e intransferível" (PINTO, 2005, P. 59). Essas necessidades surgem do confronto do organismo com o ambiente, ao longo do qual são gerados os problemas. O computador é desprovido de problemas, porquanto sua própria existência é a solução para um problema humano – a intencionalidade maquínica é de terceira pessoa.

Diante dos problemas, o homem cria representações mentais das possíveis opções, dentre as quais, em uma operação mental subsequente, escolhe alguma. Essa escolha se pauta por uma finalidade auto impingida pelo ser humano, com sua ideação abstrata. A essência do comportamento inteligente está nessa capacidade de auto definição do propósito, e não tanto na criação das opções. A máquina computacional ingressa no plano de resolução de problemas apenas pela mão de seu construtor, humano, que age em função dos interesses de sua existência em um determinado momento do processo histórico. A capacidade de a máquina fazer escolhas, tomar iniciativas e fazer outras imitações do comportamento inteligente resume-se a uma transferência de poderes, na qual o cérebro humano, único órgão capaz de elaborar projetos, concebe um projeto especial, o de uma máquina elaboradora de projetos. Pinto argumenta nesse sentido, afirmando:

> os órgãos artificiais reguladores são efetivamente o próprio sistema nervoso do animal hominizado, manifestando-se numa capacidade elevada a um nível qualitativamente superior, pois, em vez de regular diretamente a máquina ou o aparelho, regula o regulador (PINTO, 2005, P. 124).

ARGUMENTOS FAVORÁVEIS

Turing formulou a seguinte pergunta: "não podem acaso as máquinas realizar algo que deveria ser descrito como pensamento, mas que é muito diferente do que um homem faz?" (TURING, 1996, p. 24). Estava lançada a semente para a teoria computacional da mente, claramente funcionalista, assumindo como pressuposto ser desnecessário saber como o cérebro funciona para saber como a mente funciona. Os processos mentais são processos computacionais sobre elementos formais, podendo ser realizados por meio de diferentes acionamentos cerebrais, da mesma forma que um software pode ser rodado em diferentes hardwares.

Boden argumenta que o computador pode ter uma intencionalidade de primeira pessoa, recorrendo ao exemplo simples do jogo da velha: "há, claramente, um conhecimento considerável da estratégia e das táticas – não apenas as 'regras' – envolvido, gerando as escolhas do programa não apenas em relação ao 'o que' dizer, mas também ao 'como dizer'" (BODEN, 1981, P. 186, tradução livre). Um aspecto central da ação intencional é ser guiada

por uma ideia do objetivo de uma forma flexível e inteligente. Boden prossegue afirmando que "entre os primeiros programas de Inteligência Artificial havia alguns que resolviam problemas mantendo uma ideia do objetivo firme na mente e raciocinando de volta sobre si mesmos" (BODEN, 1981, p. 269).

Contra a possibilidade de uma intencionalidade de primeira pessoa aplicada a máquinas computacionais usualmente se levanta o Teorema de Gödel, referindo-se à Proposição VI do referido autor: "Proposition VI: To every ω-consistent recursive class c of formulae there correspond recursive class-signs r, such that neither υ Gen r nor Neg (υ Gen r) belongs to Flg (c) (where υ is the free variable of r)" (GÖDEL, 1992, p. 57). O próprio Gödel, explicando o que buscou provar, afirma que se trata do fato de que problemas relativamente simples na teoria dos números ordinais inteiros não podem ser decididos a partir de seus axiomas. Ou seja, existem proposições que não podem ser provadas ou *descomprovadas* dentro do sistema. A prova gödeliana, portanto, não se relaciona diretamente à questão de se os computadores poderão ou não pensar. Sua relevância consiste no fato de apontar para a existência de limites nos sistemas formais (inclusive para a lógica formal).

Como a programação de computadores opera basicamente com a lógica booleana, a prova de Gödel coloca limites intransponíveis para aquilo que um computador poderá fazer, inclusive quanto à sua suposta capacidade cognitiva. Porém, a réplica dos fundadores da IA, como Turing, consiste em afirmar a irrelevância desse elemento, uma vez que também existiriam limites para a capacidade cognitiva humana:

> a resposta mais simples a esse argumento é a de que, embora esteja estabelecido que há limitações aos poderes de qualquer máquina específica, enunciou-se apenas, sem qualquer espécie de prova, que nenhuma limitação desse tipo se aplica ao intelecto humano (TURING, 1996, P. 38).

Aqui vale recuperar também uma argumentação de Hofstadter:

> ocorre que nenhum método algoritmo pode dizer como aplicar método de Gödel a todos os tipos possíveis de sistemas formais. E, a menos que se tenha inclinações algo místicas, tem-se de concluir, portanto, que qualquer ser humano simplesmente alcançará os limites de sua própria capacidade de gödelização em algum ponto (HOFSTADTER, 2001, p. 21).

Outra linha de argumentação é a de que os computadores não são previsivelmente determinísticos, associada à uma contrapartida de que a imprevisibilidade da ação humana é geralmente exagerada. Ainda que se concorde com o fato de que tudo que a máquina faz é feito segundo instruções especificadas a priori, não se pode afirmar que o programador seja capaz de antever tudo o que a máquina vá fazer, nem que o programa vá fazer tudo e apenas aquilo que o programador pretendia que ele fizesse. Hofstadter argumenta que a complexidade introduz diferenças qualitativas, acarretando que, a partir de certo nível de complexidade, a máquina deixa de ser previsível:

> ela começaria a ter uma mente própria quando já não fosse totalmente previsível e inteiramente dócil, mas fosse capaz de fazer coisas que reconhecêssemos como inteligentes – não apenas cometer erros e atuar a esmo – e que não tivessem sido programadas nela (HOFSTADTER, 2001, p. 394).

Sem a pretensão de ir tão longe quanto Hofstadter, Floridi identifica modos de fazer com que um agente artificial lide com a incerteza e aprenda a partir de suas observações:

> representando o estado de conhecimento de um robô como uma distribuição probabilística sobre um conjunto de proposições atômicas, nós podemos representar a incerteza, e ao atualizarmos essas distribuições em resposta à evidências, usando o famoso teorema de Bayes, nós podemos modelar o aprendizado de um agente racional (FLORIDI, 2002, p. 161, ,tradução livre).

Outra resposta possível à questão da imprevisibilidade originária do comportamento humano e a suposta previsibilidade total das máquinas computacionais segue a linha dos pioneiros da IA: reconhecer o problema, mas devolvê-lo como um problema igualmente comum ao gênero humano. Uma percepção determinista (previsibilidade máxima) implica que os processos conscientes de vontade excluem qualquer adaptação a uma novidade genuína. Como são baseados em conhecimento causal, eles se adaptam a situações em que a ação necessária pode ser deduzida do que se passou anteriormente. Com isso, o futuro está, de certa maneira, incluído no passado, restando impossibilitado de ser totalmente novo ou imprevisto. Aceita essa linha de argumentação, os sistemas autopoiéticos, do tipo dos descritos por Floridi, saem em vantagem. As suas propriedades auto organizativas se fundam sobre o processo de utilização da desordem e do aleatório, estando, portanto, perfeitamente adaptados à verdadeira novidade, pois o aleatório é, por definição, a própria novidade. A

autopoiese seria um processo de criação e estabilização da novidade e, como tal, não seria passível de predição, tampouco poderia resultar da consciência.

Computadores podem ter interesses conflitantes, o que é imprescindível para se pensar no desenvolvimento de bases para julgamentos morais. Boden (BODEN, 1981, p. 70) também sinaliza que interesses conflitantes podem ser importantes para o desenvolvimento do sentido de propriocepção, quando identifica casos de paralisia histérica em robôs, causados por conflitos entre programas que comandam o início do movimento do membro robótico e programas centrais de controle geral do robô.

Computadores são capazes de mudar sua programação aleatoriamente, como no caso de alguns algoritmos genéticos, que geram estruturas que não poderiam ter sido geradas por versões prévias do programa. O comportamento adaptativo em sistemas informacionais, por exemplo, revela como uma comunidade de processos concorrentes se comporta como um sistema ecológico, com suas interações, estratégias e competição por recursos. Um robô desenvolvido recentemente apresentou a capacidade de aprender a mancar sozinho, após uma de suas pernas ser encurtada pelos pesquisadores. A máquina, de quatro pernas, foi equipada com vários tipos de sensores, que conseguiram criar um modelo corpóreo e, por meio de um algoritmo, corrigir o movimento com base na informação da perna encurtada. Turing já apresentava, em termos teóricos, a possibilidade de alteração endógena da programação de um computador, ao tratar da discriminação, que, tecnicamente, é a decisão que a máquina toma sobre o que fazer a seguir. Turing argumentou que essa decisão é tomada apenas parcialmente com base nos dados disponibilizados pelo programador, incorporando, para além desses, os próprios resultados da máquina. Turing exemplificou esse ponto do seguinte modo:

> Outra idéia importante é a de construir uma instrução e então obedecê-la. Isso pode ser usado, entre outras coisas, para discriminação. No exemplo que acabei de apresentar, nós podemos calcular uma quantidade que era 1 se $|1 - au|$ fosse menor que 2^{-31} ou 0. Adicionando essa quantidade à instrução que é obedecida no ponto de decisão, aquela instrução pode ser completamente alterada em seus efeitos quando $1 - au$ for finalmente reduzido a dimensões suficientemente pequenas (TURING, 2004, p. 389, tradução livre).

Haugeland defende que o funcionamento baseado em algoritmos – regras explícitas que determinam o próximo passo da máquina a cada rodada – não impede que se possa ter

heurística. Depende da forma como o resultado desejado for especificado. Para esse autor, os sistemas formais podem ter duas vidas: sintáticas, nas quais são marcadores desprovidos de significado que se movem de acordo com as regras de algum jogo auto-contido; ou semânticas, quando o sistema é interpretado e seus símbolos passam a ter relações significativas com o mundo externo. Ainda segundo Haugeland,

> Um sistema formal automático com uma interpretação tal que a semântica tome conta de si mesma é o que Daniel Dennett (1981) chamou de um engenho semântico. A descoberta de que engenhos semânticos são possíveis – que com o tipo correto de sistema formal e interpretação, uma máquina pode lidar com significados – é a inspiração básica das ciências cognitivas e da inteligência artificial (HAUGELAND, 2000, p. 45).

Aplicações de redes neurais e sistemas computacionais com características autopoiéticas emulam a capacidade humana de reconhecimento de padrões, sendo efetivamente utilizadas para analisar problemas complexos. As características autopoiéticas relacionam-se à capacidade do sistema de se auto ajustar, independentemente de seus artífices. Esse auto ajustamento representa uma atitude de controle de segunda ordem, um passo fundamental para a superação da sintática e a obtenção de engenhos semânticos genuínos. Maturana é um dos que enxergam essa possibilidade:

> poderemos na verdade projetar sistemas artificiais que experenciam a autoconsciência e a consciência, se nós os construirmos com uma estrutura plástica e um domínio de interações no qual eles possam penetrar em coordenações recursivas de coordenações de conduta (MATURANA, 1997, P. 240).

Babbage já falava no seu *Engenho Analítico* como um *engenho comendo a própria cauda,* ao demonstrar que os resultados em uma tábua podiam afetar outras colunas, alterando, dessa forma, as condições sob as quais a máquina estava operando. Com base em suas reflexões, Babbage reivindicava que sua máquina detinha a capacidade de operar segundo instruções que não tinham sido pré-programadas. O *Engenho Analítico* de Babbage era puramente mecânico, ou restrito ao nível do hardware. Contemporaneamente, algumas experiências interessantes demonstram a possibilidade de evolução no nível mecânico dos computadores, um campo de pesquisa que recebeu a designação de *hardwares evolucionários*. Novas tecnologias, como as FPGAs – *Field Programmable Gate Arrays* permitem que se obtenha evolução de circuitos no computador. As FPGAs são capazes de se reconfigurarem para agir

como qualquer circuito por meio da aplicação de sinais elétricos, o que permite que, ao invés de se fabricar um novo chip, um FPGA seja instantaneamente reconfigurado para se transformar nesse novo chip. Em um experimento que se tornou notório, um pesquisador determinou ao computador: "eu quero um chip que faça X" e deixou a FPGA se reconfigurar livremente. O resultado foi que a FPGA passou a usar minúsculos componentes para controlar o fluxo da eletricidade dentro dos circuitos, com um período de tempo inimaginavelmente pequeno em que o componente está passando de ligado para desligado ou vice-versa. O FPGA operou com estágios intermediários entre 0 e 1, algo que os engenheiros eletrônicos humanos ainda não descobriram como realizar.

Um sistema inteligente tem de ser dotado de uma lista de verdades essenciais e um conjunto de regras para deduzir suas implicações. Em sua dinâmica de funcionamento, precisa situar os objetos em categorias, de modo a poder aplicar ao novo objeto que tiver diante de si o conhecimento que adquiriu sobre objetos semelhantes no passado. Do contrário, caso tratasse cada novo objeto como uma entidade única, o sistema teria de ser entupido com os infinitos fatos/objetos do universo.

Ao ser feita, no presente, a programação se liga aos conhecimentos atualmente existentes. Sua gradativa realização introduz variações entre os elementos da realidade que, por serem infinitos, não podem estar contidos em nenhum projeto específico. Ao se cumprir, a programação se converte em fator perturbador dela mesma. Vale concluir essa linha de argumentação com uma citação de Turing:

> vamos supor que tenhamos programado uma máquina com algumas tábuas de instruções iniciais, construídas de tal forma que essas tabelas possam, ocasionalmente, se aparecer uma boa razão, modificarem aquelas tabelas. Alguém pode imaginar que, após a máquina operar por algum tempo, as instruções teriam se alterado tanto que não seriam reconhecíveis, mas, apesar disso esse alguém teria de admitir ainda ser aquela máquina que ainda estava fazendo cálculos muito significativos. Possivelmente, a máquina pode estar ainda gerando resultados do tipo desejado quando foi inicialmente programada, mas de uma forma muito mais eficiente. Em tal situação, esse alguém teria de admitir que o progresso da máquina não foi antevisto quando suas instruções originais foram alimentadas (TURING, 2004, p. 393, tradução livre).

No âmbito dessa discussão, há os que chamam a atenção para o fato de que se está em tela a possibilidade de existência de uma IA e não de um ser humano artificial. Ser humano e ser inteligente são coisas distintas e não faz sentido entender que uma máquina, para ser

inteligente, precise ter necessidades sexuais, fome, pulso, emoções ou, ainda, um corpo com conformação humana. Parte desse problema é devido à tradição cultural, fortemente enraizada, de considerar que o que nos distingue das demais espécies é nossa capacidade de raciocinar. Com essa perspectiva, Hofstadter postula:

> talvez estejamos inconscientemente assoberbados com um chauvinismo semelhante com respeito à inteligência e, em conseqüência, com respeito ao significado. Em nosso chauvinismo, consideraríamos 'inteligente' qualquer ser com um cérebro suficientemente parecido com o nosso e recusar-nos-íamos a reconhecer como inteligente outros tipos de objetos (HOFSTADTER, 2001, p. 186).

O que está em jogo é a definição de quais seriam os predicados que estamos dispostos a atribuir às máquinas, sem que essa atribuição resulte em uma ontologia ingênua e eticamente inerte.

A aceitação da equivalência entre o ser o fazer dos funcionalistas obstrui a argumentação de que algo que se comporta conscientemente não seja consciente[6]. Adversamente, ainda que se repila a equivalência ser/fazer, há que se reconhecer seus iso-resultados. Nos dizeres de Wittgenstein,

> que haja uma regra geral por meio da qual o músico pode extrair a sinfonia da partitura, uma por meio da qual se pode derivar a sinfonia dos sulcos do disco e, segundo a primeira regra, derivar novamente a partitura, é precisamente nisso que consiste a semelhança interna dessas configurações, que parecem tão completamente diferentes (WITTGENSTEIN, 2001, p. 167).

Como seres humanos, podemos aprender a imitar as Máquinas de Turing. Logo, por definição, somos *no mínimo* Máquinas de Turing.

---

[6] Teixeira revela sua preocupação com a aceitação integral da perspectiva funcionalista: "esse salto corresponderia também a alguma quintessência que, segundo Descartes, ficaria faltando na forma de um autômato, pois, na medida em que ser consciente não seria uma propriedade física, a replicação física integral de um cérebro não implicaria, necessariamente, na replicação do caráter consciente dos estados mentais que esse autômato poderia vir a ter" (TEIXEIRA, 2000, p. 77). Teixeira parece se referir à afirmação cartesiana sobre os autômatos: "primeiro, eles não podem jamais usar palavras ou outros sinais construídos, como nós usamos para declarar nossos pensamentos aos outros (...) segundo, enquanto eles podem fazer muitas coisas tão bem quanto qualquer um de nós ou até melhor, eles vão infalivelmente falhar em outras, revelando que eles não agem com base em conhecimento mas apenas com base na disposição de seus órgãos" (DESCARTES, 2000, p. 20) .

ESTADOS OBJETAIS NÃO OBJETIVÁVEIS

Há uma diferença entre software como performance e software como texto. A programação de um computador não é uma ciência exata, sendo impossível, a priori, deduzir todas as consequências da execução de um programa, seja qual for o ambiente. Fetzer distingue "programas-como-textos (não carregados) e programas-como-causas (carregados), onde a verificação (humana) envolve a aplicação de métodos dedutivos a programas-como-textos" (FETZER, 2000, p. 267, tradução livre). Prosseguindo em sua argumentação, Fetzer afirma que

> provas matemáticas, teorias científicas e programas de computador qualificam-se como entidades sintáticas, mas teorias científicas e programas de computador têm uma significância semântica (para o mundo físico) que provas (na matemática pura) não possuem (FETZER, 2000, p. 268, tradução livre).

No centro de sua argumentação está a premissa de que existe uma diferença entre algoritmos como uma solução efetiva de uma tarefa e programas de computador como modelos causais desses algoritmos. Os algoritmos, logicamente especificados e formalizados, são independentes de contextos, podendo ser aplicados para a derivação de conclusões a partir de premissas sem qualquer preocupação com o propósito dos argumentos relacionados. Os programas em execução exercem influências causais sobre computadores, perdendo sua isenção relativa a contextos. As máquinas informacionais, quando operam propriamente, não são apenas circunscritas às suas instruções (*law abiding*). Elas são incorporações das instruções. Na percepção de Weizenbaum, "uma teoria escrita na forma de um programa de computador é tanto uma teoria quanto um modelo ao qual a teoria se aplica, quando colocada em um computador em execução" (WEIZENBAUM, 1976, p. 74, tradução livre).

Esse fenômeno se torna mais evidente pelo fato de que os programas atuais são escritos em linguagens de nível mais alto (como Pascal, LISP, etc.), nas quais existe uma relação do tipo um-para-muitos entre os comandos do programa e as instruções executadas pela máquina. Na linguagem de máquina, diferentemente, há algo próximo a uma relação um-para-um entre comandos e instruções executadas. Os programas atuais são escritos para máquinas virtuais, que podem ter ou não contrapartes físicas. Clark fala em programas parciais,

> uma especificação genuína que, apesar disso, cede uma boa parte do trabalho e do processo decisório a outras partes da matriz causal. Nesse sentido, é muito como um programa ordinário de computador (escrito, por exemplo, em LISP) que não especifica como ou quando alcançar certos sub-objetivos, deixando essas tarefas para dispositivos previamente incorporados ao sistema operacional (CLARK, 1998, P. 157, tradução livre).

Em um programa complexo, há várias sub-rotinas, que podem ter acesso diferencial às operações umas das outras, tanto em termos de informação sobre as ações e os efeitos dessas operações, quanto em termos de interferências possíveis nas ações de outras sub-rotinas, seja para ajudar, seja para interromper. Como um dispositivo informacional (processador simbólico), o computador transcende sua natureza originária de autômato de estados finitos.

Ao definir a máquina que veio a ter seu nome, Turing afirmou que

> o comportamento possível da máquina em qualquer momento é determinado pela m-configuração, qn e o símbolo escaneado s(r). Esse par qn,s(r) vai ser chamado 'configuração': assim a configuração determina o comportamento possível da máquina (TURING, 2004, p. 59, tradução livre).

Definida teoricamente, a Máquina de Turing pode ser instanciada tanto fisicamente quanto virtualmente. O que a Máquina de Turing fará depende do seu estoque de representações (inclusive a de si mesma e suas competências) e da forma com que as distintas representações são comparadas ou transformadas umas nas outras. Enquanto software (máquina virtual), a Máquina de Turing tem apenas representações e inferências. Instanciada em um hardware, produz causas físicas:

> um programa em execução é uma máquina de um certo tipo, uma máquina *informacional*. O texto do programa – as palavras e os símbolos que o programador compõe, que 'dizem ao computador o que fazer' – é uma máquina informacional *incorpórea*. O seu computador provê um *corpo* (GELERNTER, 1993, p. 19).

A tradução entre os comandos (semântica?) para instruções (sintáticas?) gera causas físicas (oscilações de corrente, por exemplo) no hardware. Por analogia, a semântica cerebral (vontade consciente de levantar o braço) se traduz em sintática neuromuscular (impulsos neurais enviados às fibras musculares). Resultados de alto nível (semânticos) podem ser

obtidos a partir de sintáticas diversas. As estruturas físicas, entretanto, restringem as ações e interpretações. Dada a complexidade de um computador digital atual, as dificuldades para se identificar os correlatos neurais aos estados mentais, em humanos, não são menos complicadas do que aquelas para se identificar as relações entre os estados abstratos de uma Máquina de Turing e os estados estruturais do dispositivo que os estejam implementando.

A descrição lógica de uma Máquina de Turing não inclui qualquer especificação quanto à sua natureza física, nem quanto a de seus estados. A Máquina de Turing é uma máquina abstrata, que pode ser fisicamente realizada em praticamente qualquer tipo de substância. Conforme argumentado por Teixeira, Máquinas de Turing "podem ser construídas com qualquer tipo de material, até com pedacinhos de papel e latas de cerveja vazias. O que importa é a realização de uma função seja por que meio for" (TEIXEIRA, 2004, p. 88). Porém, instanciar uma Máquina de Turing com esses materiais e instanciar em hardware de computador apropriado causa resultados diferentes. A instância é a atualização do programa. O programa é para a instância o que a língua é para o ato de fala.

No primeiro computador digital, ENIAC, a programação era física e um programa típico envolvia milhares de cabos, conectados à mão, ponto a ponto, em grandes tábuas de programação. Por analogia, essa primeira versão digital da Máquina de Turing universal não era muito diferente dos pedacinhos de papel e latas de cerveja. Até os dias atuais, as operações mais comuns – adição, subtração, multiplicação – já estão inscritas na máquina, ou seja, os circuitos impressos são arranjados de tal forma que efetuam automaticamente a operação desejada. Notoriamente, os avanços mais retumbantes da IA, como os programas vencedores de xadrez, envolvem utilização de hardware especializado. Quanto mais especializada a máquina, mais sua arquitetura física reflete a estrutura de suas computações. Em uma máquina de finalidades gerais, a correspondência entre forma e função é mais fraca, e a estrutura instantânea da computação é determinada pelos detalhes do programa em execução. No nível do ENIAC, *comando* era igual a *instrução*. Em níveis superiores, o máximo que se pode afirmar é que há uma *token identity* entre comandos e instruções, similar à *token identity* entre qualia e assembléias de neurônios.

Correndo o risco de empobrecer a argumentação, exemplificamos o que pretendemos com dois computadores similares, rodando o mesmo programa, um tem uma interrupção, o

outro não. O que queremos afirmar é que, quando instanciadas em uma base física, os estados possíveis da máquina passam a ser determinados por [qn,s(r) + base física]. E que essa base física, quando tratamos dos modernos computadores digitais, pode gerar resultados iguais a partir de entradas (inputs) diferentes. Ou seja, um processador de texto em execução em máquinas similares pode gerar o mesmo resultado, apesar de seu conjunto de circuitos integrados estarem em situações físicas diferentes.

A situação física real dos computadores é inacessível a outra máquina, de modo análogo à forma como os qualia de um homem são inacessíveis aos outros homens. Destaque-se a vasta quantidade de fenômenos paralelos intercorrentes em um computador em funcionamento, como, por exemplo, o fato de que a produção de calor em resistências ôhmicas de computadores digitais faz com que essas resistências mudem, em uma percentagem mínima, seu valor. Portanto, para acessar a exata situação física da outra máquina, um computador teria que *ser* a outra máquina. Pretendemos aqui uma argumentação similar à de Thomas Nagel, em seu artigo clássico *What is like to be a bat?*, embora não compartilhemos de suas pretensões dualistas.

O que se afirma é que uma Máquina de Turing instanciada em uma base física qualquer passa a ser determinada por [qn,s(r) + base física] e esse estado é único (momentum) e irreprodutível – um estado objetal não passível de objetivação. Esses estados estariam na base uma possível derivação de traços de singularidade (talvez até personalidade) em máquinas.

Argumenta-se que uma Máquina de Turing é determinística porque cada novo estado é exclusivamente determinado por um único evento de entrada. Porém, o mesmo pode ser alegado para os componentes mais básicos dos seres vivos, como as células, cujo comportamento pode ser calculado por uma função recorrente geral, em qualquer grau de precisão desejado, desde que exista uma descrição suficientemente precisa do estado interno da célula e do meio circundante. Em um nível ainda mais elementar, o das moléculas de DNA, essa precisão de comportamento é ainda mais absoluta e determinística[7].

[7] Essa linha de argumentação encontra sustentação em outros autores. Norbert Wiener afirmou que "os seres vivos não são vivos além do nível das moléculas" WIENER (1979, p. 52) . Teixeira postulou que "as relações entre o vivo e o não-vivo são cada vez mais promíscuas, o que põe em risco o argumento que Leibniz usava para nos separar dos autônomos" (TEIXEIRA, 2006, p. 51).

O que está no cerne da argumentação que propugna o nível inultrapassável de determinismo da máquina é, na verdade, a defesa de que essa nunca poderá experimentar *Empfindungen* – sentimentos e experiências em estado bruto – ou, para usar uma terminologia mais comum no campo da Filosofia da Mente, os computadores nunca poderão ter qualia. Contudo, não existem argumentos que defendam que uma célula tenha qualia, muito menos uma molécula de DNA. Do mesmo modo que os processos conscientes de alto nível (entre eles os qualia) são experimentados de forma independente de um conhecimento funcional dos processos de nível mais baixo nos quais se sustentam (transações neuronais), um programa de computador incorpora inúmeros subprogramas. O resultado de alto nível com o qual se interage (a tela que se vê quando se trabalha com um processador de textos, por exemplo) independe de um conhecimento preciso sobre como as sub-rotinas de nível mais baixo estão realizando seu trabalho.

Pode-se argumentar, ainda, que a camada digital é aplicada sobre uma camada estritamente física e que a partir dessa aplicação acaba-se o espaço para qualquer possibilidade de qualia. No nível elétrico do microprocessador, voltagens superiores a 3,8V são traduzidas como *uns* e voltagens inferiores como *zeros* e, a partir daí, o comportamento da máquina digital seria completamente determinístico. Ocorre que a máquina não se reduz ao microprocessador e os computadores modernos são verdadeiros complexos de componentes, interagindo de maneira dinâmica, e gerando possibilidades para resultados diferentes.

A simulação de comportamento inteligente deve ter como ponto de partida os comportamentos simples, rotineiros, que não carecem da existência prévia de representações. Seria uma entidade situada fisicamente (o que significa abrir mão de construir um modelo completo do meio-ambiente para então agir sobre ele) e corporificada (capaz de distinguir verbos e substantivos). A inteligência surgiria nas interações dessa entidade com o mundo, na medida em que tiver de resolver problemas (como, segundo a perspectiva evolucionista, acontece na natureza). Dada essa característica emergente da inteligência, ela não precisa ser pré-programada.

Os estados objetais não objetiváveis poderiam também justificar a falibilidade das máquinas computacionais e, portanto, seu passo decisivo rumo à inteligência. Turing parece ter antevisto essa possibilidade ao afirmar que "o argumento de Gödel e outros teoremas se

apóiam essencialmente na condição de que a máquina não cometa erros. Mas isso não é um requisito para a inteligência" (TURING, 2004, p. 211, tradução livre). No entender de Turing, o que o teorema de Gödel e outros resultados correlatos (como a própria Tese Church-Turing) demonstram é que se forem utilizadas Máquinas de Turing Universais para propósitos como o de determinar a verdade ou a falsidade de teoremas matemáticos, e não houver tolerância para a eventualidade de um resultado errado, nenhuma máquina será capaz, em alguns casos, de chegar a uma resposta. Segundo Teixeira (2004), essa incapacidade de se chegar a uma resposta não significa, necessariamente, que se está diante de uma situação de não-algoritmicidade ou de incomputabilidade[8]. Pode-se estar diante de um problema transcomputável: "um problema transcomputável é um problema intratável cujo procedimento algorítmico de solução não pode ser obtido em tempo eficiente a despeito de qualquer aperfeiçoamento do hardware do computador utilizado" (TEIXEIRA, 2004, p. 99).

Continuando sua linha de argumentação, Turing afirma que

> se se espera que uma máquina seja infalível, ela não pode ser inteligente. Há vários teoremas matemáticos que dizem exatamente isto. Mas esses teoremas não dizem nada a respeito de quanta inteligência pode ser demonstrada se uma máquina não tiver qualquer pretensão de infalibilidade (TURING, 2004, p. 394).

Curiosamente, na revanche em que derrotou *Deep Blue*, Kasparov afirmou ter feito lances ruins de propósito e ter jogado aquém de suas capacidades. A inteligência estaria na capacidade de errar? Ou seria o erro a forma de burlar a sintática formal dos programas-como-texto e ingressar no mundo da semântica?

CONSIDERAÇÕES FINAIS

Os últimos desenvolvimentos da IA trouxeram expectativas do surgimento de uma nova entidade, capaz de comportamentos autônomos e inteligentes. Foram apresentados argumentoss favoráveis e contrárias, deixando a questão inconclusa, buscando manter um equilíbrio entre o tecnoentusiasmo e o tecnocatastrofismo.

[8] Teixeira (2004, p. 92) registra que "se pudermos saber se existe ou não uma outra máquina de Turing que nos permita saber se uma máquina de Turing pára ou não, teremos encontrado o procedimento mecânico algorítmico cuja possibilidade de existência Hilbert questionava".

Entretanto, o artigo deixa registrada uma vereda sobre possibilidades de singularização das máquinas digitais, a partir da discussão dos estados objetais não objetiváveis. A instanciação material da Máquina de Turing (conceito abstrato) traria a possibilidade, teórica, de construtos identitários singulares a máquinas digitais. A existência dessa possibilidade deixa aberto o caminho para que as máquinas digitais se tornem autônomas e inteligentes, embora isso não signifique que venham a se tornar humanas. É provável que se essa máquina um dia vier a pensar, ao tentarmos ler esses pensamentos, estejamos como quando nos deparamos com o leão de Wittgenstein.

Fazendo um paralelo, persiste até os nossos dias o dilema mente-cérebro, para alguns o último mistério verdadeiramente filosófico. Mas a insolubilidade do problema mente-cérebro não nos impediu de existirmos e de termos pensamentos e ações autônomas. O fato de não conseguirmos formular um modelo científico para que ocorra o surgimento de um computador 'pensante' não deve ser considerado um óbice definitivo a essa possibilidade. Trazendo o argumento a planos mais modestos, o que pretendemos sinalizar é que a 'singularidade' e 'autonomia' podem vir a ser contrafacções maquínicas de 'individualidade' e 'inteligência'.

A imersão em um espaço crescentemente povoado por elementos imateriais remete à realidade virtual, que se apresenta como promessa de liberdade absoluta, espaço de domínio total, quase mágico, no qual uma palavra ou um gesto podem mudar tudo. Qualquer coisa passível de ser construída semioticamente pode acontecer na realidade virtual. Purificado das mazelas do espaço natural, o espaço virtual será tudo aquilo que o homem pretenda fazer com ele – espaço aberto às pretendidas extensões de nossas capacidades de percepção e ação. Mas, por trás dessa máxima liberdade, encontra-se o indivíduo castrado, não mais *produtor* de seu espaço, apenas *usuário* de um mundo programado por terceiros.

O desenvolvimento da IA traz ao cenário uma possibilidade angustiante: que de vicários – *agem em nosso lugar* – passem a vampiros – *suguem nossas energias*. Todo artefato gera dependência (apropriamo-nos deles e passamos a os considerar como parte de nossos organismos). E dependência sempre gera vulnerabilidade. Quanto mais 'agente' a IA for em nosso lugar, menos 'agentes' seremos nós mesmos, restando passivos e obseqüentes.

Trata-se de uma reflexão, ainda muito embrionária, sobre os efeitos existenciais e psicológicos da possibilidade de virmos a conviver com novas entidades autônomas e

singulares, criadas por nós mesmos, mas capazes de nos superar. Possibilidade que se afigura assustadora. Por outro lado, um ambiente povoado por tais entidades será um ambiente distinto, exercendo condicionamentos diferentes sobre nós.

REFERÊNCIAS

BEAVERS, Anthony F. Phenomenology and artificial intelligence. In: MOOR, James H.; BYNUM, Terrell Ward (Org.). **Cyberphilosophy:** the intersection of computing and philosophy. Blackwell Publishing Ltd., Oxford, UK, 2002. 14p.

BODEN, Margaret A. **Minds and mechanisms:** philosophical psychology and computational models. New York: Cornell University Press, 1981.

CAPUCCI, Píer Luigi. Por uma arte do futuro. In: DOMINGUES, Diana (Org.). **A arte no século XXI:** a humanização das tecnologias. São Paulo: Fundação Editora da UNESP, 1997. 16p.

CHURCHLAND, Paul M. **Matéria e consciência:** uma introdução contemporânea à filosofia da mente. São Paulo: Editora UNESP, 2004.

CLARK, Andy. **Being there:** putting brain, body, and world together again. Cambridge, Massachusetts: The MIT Press, 1998.

DENNETT, Daniel C. **Consciousness explained.** New York: Back Bay Books, 1991.

DESCARTES, René. **As paixões da alma**. São Paulo: Martins Fontes, 1649/1998.

DREYFUS, Hubert L. Desmistificador da inteligência artificial, ante Edward Feigenbaum, especialista em sistemas especializados. In: PESSIS-PASTERNAK, Guitta. **Do caos à inteligência artificial**: quando os cientistas se interrogam. São Paulo: Editora da Universidade Estadual Paulista, 1993. 8p.

FETZER, James H. Philosophy and computer science: reflections on the program verification debate. In: BYNUM, Terrell Ward; MOOR, James H. (Org.). **The digital phoenix:** how computers are changing philosophy. Blackwell Publishers: Oxford, UK, 2000. 20p.

FLORIDI, Luciano. **Philosophy and computing:** an introduction. New York: Routledge, 1999.

_____. What is the philosophy of information? In: MOOR, James H.; BYNUM, Terrell Ward. **Cyberphilosophy:** the intersection of computing and philosophy. Blackwell Publishing Ltd., Oxford, UK, 2002. 32p.

GELERNTER, David. **Mirror worlds or the day software puts the universe in a shoebox...how it will happen and what it will mean**. New York: Oxford University Press, 1992.

GÖDEL, Kurt. **On formally undecidable propositions of Principia Mathematica and related systems.** New York: Dover Publications Inc., 1992.

GORZ, André. **O imaterial:** conhecimento, valor e capital. São Paulo: Annablume, 2005.

HAUGELAND, J. Semantic engines: an introduction to mind design. In: CUMMINS, Robert; CUMMINS, Denise Dellarosa (Org.). **Minds, Brains and Computers:** the foundations of cognitive science, an anthology. Malden, USA: Blackwell Publishers Inc., 2000. 21p.

HEIDEGGER, Martin. **Que é uma coisa?** Doutrina de Kant dos princípios transcendentais. Lisboa: Edições 70, 1987.

HOFSTADTER, Douglas R. **Gödel, Escher, Bach:** um entrelaçamento de gênios brilhantes. Brasília: Editora Universidade de Brasília; São Paulo: Imprensa Oficial do Estado, 2001.

LÉVY, Pierre. . **A máquina universo:** criação, cognição e cultura informática. Porto Alegre: Artmed, 1998.

_____. **Cibercultura**. São Paulo: Ed. 34, 1999.

MANER, Walter. Heuristic methods for computer ethics. In: MOOR, James H.; BYNUM, Terrell Ward. **Cyberphilosophy:** the intersection of computing and philosophy. Blackwell Publishing Ltd., Oxford, UK, 2002. 22p.

MATURANA, Humberto. **A ontologia da realidade**. Belo Horizonte: Ed. UFMG, 1997.

NORMAN, Donald A. **Things that make us smart:** defending human attributes in the age of the machine. New York: Addison-Wesley Publishing Company, 1993.

PINTO, Álvaro Vieira. **O conceito de tecnologia**, Rio de Janeiro: Contraponto, 2005. v. 2.

POLLOCK, John L. Procedural epistemology. In: BYNUM, Terrell Ward; MOOR, James H. (Org.). **The digital phoenix:** how computers are changing philosophy. Blackwell Publishers: Oxford, UK, 2000. 17p.

TEIXEIRA, João de Fernandes. **Filosofia e ciência cognitiva.** Petrópolis, RJ: Vozes, 2004.

_____. **Filosofia da mente:** neurociência, cognição e comportamento. São Carlos, SP: Claraluz, 2005.

TURING, Alan. Computação e inteligência. In: TEIXEIRA, João de Fernandes. **Cérebros, máquinas e consciência**: uma introdução à filosofia da mente. São Carlos, SP: EDUFSCar, 1996. 16p.

_____. On computable numbers, with an application to the Entscheidungsproblem (1936). In: COPELAND, B. Jack (Org.). **The essential Turing:** the ideas that gave birth to the computer age. Oxford: Clarendon Press, 2004. 32p.

WEIZENBAUM, Joseph. **Computer power and human reason:** from judgment to calculation. San Francisco: W.H.Freeman and Company, 1976.

WITTGENSTEIN, Ludwig. **Tractatus Lógico-Philosophicus**. São Paulo: Editora da Universidade de São Paulo, 2001.

Recebido: 09/05/2024
Aprovado: 02/07/2024