Uma investigação acerca da possibilidade de agência em sistemas artificiais

Camila da Cruz Silva*

RESUMO

O presente artigo tem por objetivo investigar, no contexto das tecnologias informacionais emergentes, a possibilidade de sistemas artificiais serem inseridos na categoria de agentes. Para realizar o objetivo proposto, explicitaremos as condições que permitem classificar um sistema artificial enquanto agente, apresentando, assim, uma análise acerca do conceito de agência tanto em teorias da ação consideradas como padrão, que têm como ponto comum a noção de ação intencional, como em teorias alternativas da agência, que não pressupõem, necessariamente, funções cognitivas de segunda ordem. Em um segundo momento investigaremos a noção de autonomia, na medida em que esta é uma característica fundamental da atribuição não trivial de agência. Por fim, procuraremos mostrar que sistemas artificiais que forem capazes de ir além de sua programação inicial, na medida em que aprendem com a própria experiência e alteram o rumo de sua ação como resultado de tal aprendizagem, podem ser inseridos na categoria de agentes autônomos dentro de condições e contextos específicos.

Palavras-chave: Sistemas artificias. Agência. Intencionalidade. Autonomia. Agentes artificiais.

An investigation on the possibility of agency in artificial systems

ABSTRACT

The aim of this article is to investigate, in the context of emerging information technologies, the possibility of including artificial systems in the category of agents. In order to achieve this goal, we will explain the conditions that allow an artificial system to be classified as an agent, and thus present an analysis of the concept of agency both in standard theories of action, which have the notion of intentional action as their common point, and in alternative theories of agency, which do not necessarily presuppose second-order cognitive functions. Secondly, we will examine the notion of autonomy, insofar as this is a fundamental feature of the non-trivial attribution of agency. Finally, we will try to show that artificial systems that are capable of going beyond their initial programming, in the sense that they learn from their own experience and change their course of action as a result of that learning, can be included in the category of autonomous agents under certain conditions and contexts.

Keywords: Artificial systems. Agency. Intentionality. Autonomy. Artificial agents.

* Doutoranda em Filosofia pela Pontifícia Universidade Católica do Paraná (PUCPR). Possui Mestrado em Filosofia pela UNESP de Marília (2020) com auxílio da bolsa CAPES. Possui graduação (licenciatura e bacharelado) em Filosofia pela UNESP de Marília (2018). É membro do Grupo Acadêmico de Estudos Cognitivos (GAEC). Tem experiência na área de Filosofia com ênfase em Filosofia da Mente e Filosofia da Informação, interessada principalmente nos seguintes temas: Sistemas artificiais, Relação entre Filosofia e Tecnologia, Responsabilidade, Ética dos algoritmos. Email: camilamingroni1993@gmail.com. Orcid: https://orcid.org/0000-0002-8698-7509.

Kairós: Revista Acadêmica da Prainha
ISSN: 1807-5096
e-ISSN: 2357-9420
Fortaleza,
v. 20, n. 1, 2024

Introdução

Num sentindo de aperfeiçoamento e progresso, é possível observar que tanto seres humanos quanto animais buscam formas de adaptar seu ambiente de acordo com suas necessidades. Assim, após centenas de anos de aprimoramento tecnológico, que data desde a fabricação das primeiras ferramentas produzidas pelos seres humanos, nos encontramos em um contexto no qual dispositivos eletrônicos desempenham diversas funções e tarefas que antes eram reservadas apenas aos seres humanos.

Na atualidade observamos a construção e aplicação de artefatos tecnológicos altamente sofisticados e complexos, cujas funcionalidades muitas vezes escapam do controle e da compreensão não só apenas de seus usuários, mas também, muitas vezes, dos próprios idealizadores de tais tecnologias.

Tomando como ponto de partida a caracterização de nosso contexto atual a partir da computação ubíqua (WEISER, 1991), que consiste na computação incessantemente integrada aos ambientes, e muitas vezes, à Internet, e o surgimento de tecnologias de informação e comunicação também ubíquas, é possível afirmar que nós, seres humanos, passamos a resolver muitas questões cotidianas a partir de nossos celulares e computadores inteligentes, para citar apenas alguns dos exemplos mais comuns.

Neste contexto, é importante enfatizar também que, devido ao isolamento social oriundo da pandemia do COVID-19 declarada oficialmente em março de 2020[1], o uso de artefatos eletrônicos de informação digital conectados à rede de Internet teve um aumento e uma adesão sem precedentes, na medida em que as pessoas passaram a fazer a manutenção de suas relações sociais, estabelecer novas relações, executar funções de trabalho e realizar inúmeras atividades que antes exigiam a presencialidade.

Isto posto, este trabalho se situa no contexto das tecnologias de informação e comunicação digitais e do desenvolvimento e aprimoramento de certos artefatos tecnológicos. Entendemos que a evolução de tecnologias digitais em conjunto com o aperfeiçoamento da programação de tais tecnologias, na medida em que são

[1] Disponível em: https://www.unasus.gov.br/noticia/organizacao-mundial-de-saude-declara-pandemia-de-coronavirus.

mediados pela computação ubíqua, tem como objetivo construir tecnologias que integrem em sua constituição elevados graus de automação que possibilitem ações sem a intervenção direta dos criadores dessas tecnologias.

Nossa proposta consiste, assim, em investigar a possibilidade de sistemas artificiais, concebidos nesse sofisticado ambiente tecnológico-informacional, serem incluídos na categoria de agentes. Para tanto, faremos uma análise dos conceitos de agência em diferentes contextos e do conceito de autonomia para verificar se existem condições favoráveis que permitiriam tal inclusão.

Análise do conceito de agência

Para dar início ao tópico que tratará sobre a agência em sistemas naturais e artificiais, é fundamental introduzir partes essenciais do debate filosófico realizado para esclarecer o conceito de agência. O estudo desse conceito gerou diversas interpretações, considerando que, em sentido geral, conforme Schlosser (2019) indica, agências podem ser encontradas em todo lugar. “Sempre que as entidades entram em relações causais, pode-se dizer que elas agem umas sobre as outras e interagem entre si, provocando mudanças umas nas outras” (SCHLOSSER, 2019, p. 2, tradução nossa).

Desse modo, dada a complexidade em definir o conceito de agência e a possibilidade de sua admissão em diversas e distintas entidades, verificaremos a possibilidade de sistemas artificiais serem considerados como instanciadores de agência. Para tanto, enfatizamos ainda que os conceitos de *agência*, *agente* e *ação* são interligados e mutuamente dependentes na medida em que a *agência* é propriedade do *agente* que pratica ou realiza *ações* no mundo.

Segundo Markus Schlosser (2019), os estudos sobre agência foram relacionados aos estudos sobre a ação intencional. Essa forma de relacionar os dois temas pode ser encontrada em trabalhos de figuras históricas da filosofia, tais como Aristóteles e Hume, e na filosofia analítica contemporânea, os trabalhos mais influentes sobre os conceitos de ação intencional e de agência são remetidos à Elizabeth Anscombe (1957) e a Donald Davidson (1963).

Assim, a filosofia da ação contemporânea fornece uma concepção padrão da ação e uma teoria padrão da ação. Elizabeth Anscombe (1957) – responsável pela

concepção padrão da ação – analisa e interpreta a ação intencional. E Donald Davidson (1963) – responsável pela teoria padrão da ação – explica a intencionalidade da ação nos termos de uma causação decorrente de estados mentais e de eventos relacionados ao agente.

O ponto comum entre Anscombe e Davidson se encontra na consideração dos autores pela explicação da ação a partir dos termos da intencionalidade e da ação intencional. Porém, como indica Schlosser (2019, p. 3, grifos do autor, tradução nossa), os debates sobre o conceito de agência estabeleceram certa resistência e deixaram de lado sua relação com a ação intencional.

> [...] essa resistência equivale à rejeição da *concepção* padrão de ação, em alguns casos; em outros, equivale à rejeição da *teoria* padrão da ação e, em outros casos, ainda, equivale à alegação mais modesta de que existem diferentes *tipos* de agência.

Assim, antes de explorarmos as concepções alternativas de agência, analisaremos a concepção de agência considerada padrão na Filosofia da Mente e da Ação contemporâneas, apresentando suas principais características.

De forma resumida, a concepção padrão não está comprometida com a consideração feita pela teoria padrão que explica a ação intencional a partir de uma causa ligada à razão. A teoria padrão admite que uma ação poderá ser classificada como intencional e pautada por razões apenas no caso dessa ação ser causada por estados e eventos mentais adequados.

Para Donald Davidson (1963; 1971), uma ação é sempre intencional. Isso significa que a ação poderá sempre ser explicada de acordo com uma razão do ponto de vista do agente. Por sua vez, um acontecimento que não possa ser explicado ou racionalizado pelo agente será considerado como um evento.

Segundo Donald Davidson (1963), a racionalização de uma ação é uma espécie de explicação causal ordinária. Para o autor, essa racionalização só acontece se o agente permitir que outras pessoas vejam qualquer característica que o mesmo tenha visto, ou que pensou ter visto, para executar a sua ação. Essas características incluem quaisquer consequências ou aspectos que indiquem o motivo pelo qual o agente tenha executado determinada ação. Esse motivo pode ser reconhecido como um desejo do agente, ou como o processo meditativo e reflexivo geralmente realizado antes da tomada de decisões. O motivo pode ser ainda algo que beneficia ou obriga

o agente a agir de determinada maneira. Para Davidson, não é suficiente explicar o motivo de um agente agir apenas dizendo que a ação o atraiu: é necessário indicar o que houve de apelativo na ação para que ela tenha acontecido.

O que significa, então, dizer que alguém agiu por determinado motivo, ou que uma ação possui uma razão que a sustenta? De acordo com Davidson, um agente que racionaliza sua ação pode ser classificado como "[...] (a) tendo alguma *pro attitude* que direciona a ações de um certo tipo, e (b) acreditando (ou sabendo, percebendo, notando, lembrando) que sua ação é daquele tipo" (DAVIDSON, 1963, p. 685, tradução nossa).

Entendemos que *pro attitudes,* traduzidas como atitudes proativas, podem ser interpretadas como sendo atitudes direcionadoras para ações de um certo tipo, que incluem desejos, sugestões, valores morais, princípios estéticos, convenções sociais e econômicas e objetivos públicos e privados. Assim, racionalizar uma ação explicando por que um agente agiu de certa forma é, muitas vezes, uma questão de nomear uma *pro attitude*, uma crença relacionada, ou ambos (DAVIDSON, 1963).

A essa relação causal, ou seja, a relação que implica encontrar um motivo que racionalize e explique por que o agente agiu de determinada maneira, Davidson dá o nome de razão primária (*primary reason*). As razões primárias, que justificam a execução de ações, consistem, assim, em uma *pro attitude* e em uma crença relacionada da maneira correta.

Para entendermos o significado de crença e de *pro atitude,* corretamente relacionados, trazemos as considerações de Jeff Speaks (2009), que utiliza o clássico exemplo de Davidson, no qual o agente acende a luz e acaba por alertar um ladrão que se encontra nos arredores.

Neste exemplo, Davidson (1963) explica que aciona o interruptor, acende a luz e com isso seu quarto fica iluminado, porém, sem o seu conhecimento, essa ação de acender a luz tem como consequência alertar um possível invasor de que a casa não está vazia. Segundo o autor, sua ação é constituída unicamente pelo acionamento do interruptor, e as consequências intencionais dessa ação, ou seja, a luz acender e o quarto se iluminar, são as explicações que justificam tal ação. Mas essas consequências não são suficientes para racionalizar a outra consequência não intencional gerada, que foi alertar o possível invasor. Segundo Speaks:

> [...] por que dizemos que o agente ligar o interruptor foi intencional, mas que ele alertar o ladrão não foi? Parece quase impossível não dizer que a diferença é que o agente queria (ou seja, tinha uma certa atitude proativa) acender a luz, mas não queria alertar um ladrão; ou, novamente, que o agente sabia que sua ação acenderia a luz, mas não sabia que sua ação alertaria o ladrão. Reflexões sobre casos como esse parecem deixar claro que a diferença entre atos intencionais e não intencionais pode ser encontrada nas crenças e desejos do agente (SPEAKS, 2009, p. 2, tradução nossa).

É neste sentido que as razões primárias são justificativas para uma ação. Com o exemplo mencionado acima entendemos que na *pro attitude* ilustrada no ato de acionar o interruptor, e na crença de que o fazendo a luz se acenderá e iluminará o ambiente, a ação está racionalmente justificada, é uma ação intencional. Porém, como acionar o interruptor não se relaciona com alertar o invasor, essa ação não é intencional no sentido de que não existe uma crença que se alinhe ao desejo do agente.

De volta a Schlosser, é pelos motivos citados anteriormente que a teoria padrão da ação é compreendida como sendo uma teoria causal de eventos de ação intencional e, como já foi visto, a teoria padrão da ação se diferenciaria da concepção padrão da ação pelo fato da última não estar diretamente ligada a explicações do que significaria agir intencionalmente e do que significaria agir a partir de alguma racionalização da ação. Ou seja, a concepção padrão é compatível com teorias não-causais da ação intencional.

> É geralmente aceito que uma explicação da razão de uma ação geralmente torna a ação inteligível ao revelar a meta ou intenção do agente. De acordo com teorias não causais, ter objetivos ou intenções relevantes não consiste na posse de estados ou eventos mentais causalmente eficazes (SCHLOSSER, 2019, p. 6, tradução nossa).

É importante pontuar que estados mentais e eventos adequados são eventos e estados que racionalizam a ação, atribuindo-lhe uma causa/motivação a partir do ponto de vista do agente. Ou seja, esses eventos e estados mentais constituem a explicação racional que faz com que uma ação seja racionalizada, que é o mesmo que dizer que uma ação pode ser explicada e justificada.

Ainda que a ideia central da concepção padrão da ação expresse que as ações devam ser explicadas nos termos da intencionalidade e da ação intencional, tanto Davidson quanto Anscombe contribuíram para a formulação da concepção padrão da ação. Segundo Schlosser (2019, p. 4, tradução nossa), nessa formulação da ação a

“ação intencional é mais fundamental que a ação por si mesma: ações derivam e são dependentes da ação intencional”.

Porém, Davidson foi além da concepção padrão da ação ao propor que a teoria padrão da ação fornece explicações causais baseadas na razão, de forma que, para Davidson (1963), a razão opera causalmente. É por este motivo que muitas vezes essa teoria é chamada de “teoria causal da ação”, como enfatiza Schlosser (2019, p. 5-6, tradução nossa):

> Esta teoria diz, aproximadamente, que algo é uma ação intencional e que foi feito por razões apenas no caso de ter sido causado por estados mentais e eventos certos de maneira certa. Os estados mentais certos e os eventos são estados e eventos que racionalizam a ação a partir do ponto de vista do agente (tais como desejos, crenças e intenções). [...] Falando estritamente, esta é uma teoria evento-causa e consiste em uma teoria evento-causa da explicação racionalizada e em uma teoria evento-causa da ação intencional. Em conjunto com a concepção padrão, esta teoria causal nos fornece uma teoria da ação, que tem sido a teoria padrão na filosofia contemporânea da mente e da ação.

A concepção padrão, por sua vez, não estaria comprometida com explicações do que seria agir intencionalmente e a partir de razões, assim como não estaria comprometida com a natureza de explicações racionalizadas.

Elizabeth Anscombe, em seu trabalho intitulado *Intention* (1957), se propôs a clarificar o sentido do conceito ‘intenção’ e o que significa dizer que uma ação é intencional. Segundo a autora, há vários usos comuns do termo ‘intenção’. Por exemplo, quando alguém diz que “vai fazer isso e aquilo” é possível dizer que tal enunciado é uma expressão de sua ‘intenção’. Além disso, falamos sobre uma ação como sendo intencional, por exemplo, e podemos ainda perguntar qual era a intenção do agente quando praticou a ação X ou Y.

De acordo com Anscombe (1957), em cada exemplo citado acima, o conceito de ‘intenção’ é empregado com competência linguística. Porém, se tentarmos descrever este conceito considerando apenas um dos casos citados como sendo suficiente para explicá-lo em todos os demais casos, estaremos utilizando-o equivocadamente:

> [...] podemos dizer que a “Intenção sempre diz respeito ao futuro”. Mas uma ação pode ser intencional sem se preocupar de forma alguma com o futuro. Perceber isso pode nos levar a dizer que há vários sentidos da [palavra] 'intenção', e que talvez seja completamente enganoso dizer que a palavra

> "intencional" necessariamente está conectada com a palavra 'intenção', uma vez que uma ação pode ser intencional sem que contenha alguma intenção. Ou, como alternativa, podemos ficar tentados a pensar que apenas ações realizadas com certas intenções adicionais devem ser chamadas de intencionais (ANSCOMBE, 1957, p. 1, tradução nossa).

Para Elizabeth Anscombe, embora seja tentador afirmar que o conceito de intenção contenha em si diversos sentidos, é inadmissível dizer que tal conceito seja ambíguo, isto é, que possua diferentes acepções enquanto se apresenta em diferentes situações (Anscombe, 1957, p. 1). Como destaca Rachel Wiseman, em *Anscombe's Intention* (2016), a investigação de Anscombe sobre o conceito de 'intenção' se inicia a partir da pergunta: "O que distingue ações intencionais daquelas que não são?" (p. 78, tradução nossa).

A resposta a essa pergunta, segundo Anscombe (2016, p. 9), é que a ação intencional pertence a um tipo de ação em que a pergunta 'Por quê?' pode ser legitimamente colocada, isto é, é cabível perguntar a alguém "por que praticou a ação X?", e a pessoa pode oferecer uma razão para tê-la praticado.

Cabe apontar que, para Elizabeth Anscombe, dar uma razão para o agir (em resposta à pergunta "Por quê?" agi dessa ou daquela maneira) não é o mesmo que explicar causalmente uma ação, de forma que, para a autora, a ação intencional resulta de uma razão, mas esta razão não operaria causalmente, como Donald Davidson (1963) defende.

Ações intencionais são praticadas por alguma razão de forma que a resposta do agente à pergunta 'por que realizou X?' deve dar um sentido à sua ação, ou seja, o significado de uma ação intencional deve ser capaz de ser explicado através da resposta ao porquê o agente realizou X ou Y. Assim, a intenção não pode ser confundida com a causa de uma ação.

É importante pontuar que *Intention* (1957), além de ser uma obra importante para a teoria da ação estudada em Filosofia, carrega em si um significado histórico que não deve ser esquecido devido ao contexto no qual foi escrita.

Anscombe escreveu seu trabalho *Intention* para tentar explicar as *motivações* que levam um agente a praticar determinada ação. Segundo a filósofa, seu propósito original era mostrar que a concessão do título de Doutor *Honoris causa* da Universidade de Oxford ao ex-presidente dos Estados Unidos, Harry Truman, basicamente por ter lançado as bombas atômicas nas cidades japonesas de

Hiroshima e Nagasaki em 1945, seria moralmente inaceitável. Com esta decisão, Truman consentiu com a morte de mais de duzentos mil civis, a fim de defender interesses dos aliados e em nome de um bem maior: o fim da guerra e a vitória. Segundo John Haldane (2020, tradução nossa):

> Perplexa pelos defensores de Truman, [Anscombe] chegou à conclusão de que eles não compreenderam a natureza de suas ações, e foi isso que a levou a escrever *Intention*, obra na qual ela apontou que, ao fazer uma coisa (mover a mão), pode-se intencionalmente estar fazendo outra (levando à morte de seres humanos).

Nota-se que no debate acerca do conceito de agência, algumas posições ganharam mais espaço e fundaram ideias tomadas como padrão aceito para explicar a noção de agência. Porém, entendemos que tal padrão não é absoluto e não abrange todos os tipos de agência.

Um dos problemas da concepção padrão da ação é que ela está interessada em um tipo específico de agência, a agência intencional. Nesta teoria, a ação é considerada intencional quando há um conteúdo mental que pode ser representado através de proposições ou de atitudes proposicionais. Assim, a intencionalidade pode ser entendida, em seu sentido mais profundo, como a capacidade da mente de representar seus conteúdos, bem como de estar consciente dessas representações, que por fim podem ser racionalizadas a partir de conteúdos proposicionais.

Porém, parece que existem seres que são capazes de agência genuína e que não possuem estados mentais representacionais como os dos seres humanos, e quando o foco não é mais a agência humana, outros desafios emergem. Schlosser (2019, p. 16, tradução nossa) indica ainda que existem outros candidatos capazes de diferentes tipos de agência que não requerem representações mentais: “Eles incluem agência mental, agência compartilhada, agência coletiva, agência relacional e agência artificial”.

Vimos, até então, que as principais teorias aqui tratadas apresentam a intencionalidade como requisito para o reconhecimento da agência e, nesse sentido, apenas seres humanos com capacidade para justificar, reconhecer e racionalizar suas ações, seriam agentes. Essas capacidades mentais envolvem, em parte, aquilo que na filosofia da mente é compreendido como representações mentais, e para as

representações mentais acontecerem seria necessário que o agente tenha capacidade de formação de conteúdo proposicional.

Será no sentido oposto às teorias apresentadas, ao representacionismo e sua caracterização da agência intencional como o tipo padrão de agência reconhecida, que trataremos no próximo tópico sobre outros tipos de agência.

Diferentes tipos de agência

Como colocado anteriormente, parecem existir diversos tipos de agência que podem ser reconhecidos enquanto agência genuína, mas que não se encaixam nas determinações da teoria e da concepção padrão da ação. Assim, Schlosser menciona em seu trabalho mais de dez tipos de agência. Entre eles, os três tipos que o autor julga como refinados são: agência auto-controlada, agência autônoma e agência livre. Eles se diferenciam dos demais pois os outros são considerados mais básicos, ou seja, não requerem atribuição de estados mentais representacionais (SCHLOSSER, 2019).

Considerando a classificação proposta por Schlosser (2019), entendemos que, ao tratar de agência livre, agência auto-controlada e agência autônoma, o filósofo está tratando de tipos de agência que envolvem representações mentais para sua efetivação. Mais adiante em nossa investigação, trataremos tanto da autonomia de agentes naturais, quanto da autonomia de sistemas artificiais, e veremos que, ainda que um sistema artificial não possua a vida mental de um agente natural, por exemplo, de um agente humano, este também pode ser considerado como um agente autônomo em algum grau.

Os outros seis tipos de agência, de acordo com Schlosser são: agência mental, agência epistêmica, agência compartilhada, agência coletiva, agência relacional e agência artificial. É importante esclarecer primeiro que, de forma geral, os tipos diferentes de agência são postulados tendo como referência a teoria padrão da ação, de forma que se torne possível observarmos como as características de cada tipo se aproximam ou se distanciam da teoria padrão, bem como porque seria problemático admitir a teoria padrão da ação como a única aceitável para explicar todos os outros tipos de agência reconhecidos.

Segundo Schlosser (2019, p. 16), ao tratar de agência mental, pode parecer óbvio que nossas vidas mentais estão preenchidas com ações mentais na medida em que prestamos atenção, julgamos, fazemos considerações, fundamentamos, racionalizamos, aceitamos, decidimos, tentamos, deliberamos, e assim por diante. Neste sentido, se considerarmos tais casos fazendo uso da teoria padrão da ação, encontramos dois problemas:

> Primeiro, parece que tais ocorrências mentais quase nunca são, ou nunca são, ações intencionais. De acordo com a teoria padrão, um evento é uma ação intencional do tipo A apenas se o agente tiver uma intenção que inclua A em seu conteúdo. No caso básico, essa seria uma intenção de A. Em um caso instrumental, essa seria uma intenção de executar alguma outra ação B para A. [...]. Considere o pensamento de que *p*. De acordo com a teoria padrão, pensar que *p* é uma ação intencional somente se o agente tiver uma intenção que inclua "pense que *p*" em seu conteúdo (SCHLOSSER, 2019, p. 16-17, grifos do autor, tradução nossa).

Assim, o ato mental de pensar que *p* seria uma ação intencional apenas seria possível se o agente tiver uma intenção que inclua este pensamento em *p.* Esta tese é problemática uma vez que seria necessário ter outra intenção de pensar em um determinado pensamento antes de pensá-lo, podendo-se cair, inclusive, numa regressão infinita. Por sua vez, o segundo problema apontado por Schlosser (2019) se refere a casos de tomada de decisão.

> De acordo com a teoria padrão, decidir A seria uma ação intencional apenas se alguém já tivesse a intenção de tomar uma decisão que inclua "decidir A" em seu conteúdo. Isso parece, novamente, bastante estranho e problemático. Além disso, nossas razões para tomar uma decisão para A geralmente são nossas razões para A - elas são razões para executar a ação. De acordo com a teoria padrão, algo é uma ação apenas se tiver uma explicação de motivo (em termos de desejos, crenças e intenções do agente). Como as razões geralmente são motivos de ação, é novamente difícil ver como a tomada de uma decisão pode ser uma ação (2019, p. 17, tradução nossa).

Neste sentido, como razões geralmente justificam a ação, seria difícil perceber como a tomada de decisão poderia ser, por sua vez, uma ação, no mesmo sentido em que dizemos que denominamos ação, por exemplo, "cumprimentar a vizinha". Assim sendo, pensamentos dificilmente poderão ser considerados ações, pelo menos não no mesmo sentido em que se propõe o conceito de ação.

A agência epistêmica, por sua vez, diz respeito ao controle que os agentes podem exercer sobre suas crenças. Distingue-se, nesta consideração de agência,

duas posições principais: voluntarismo opinativo indireto e voluntarismo opinativo direto.

De forma resumida, o voluntarismo indireto se caracteriza a partir das maneiras segundo as quais se é possível adquirir ou revisar crenças; já o voluntarismo opinativo direto supõe ser possível ter controle voluntário direto sobre algumas crenças. Segundo Schlosser (2019, p. 18), esse controle voluntário é normalmente compreendido como o tipo de controle que os agentes exercem quando praticam ações intencionais.

Neste sentido, o desafio que se coloca a essa segunda posição é que seria preciso encontrar formações de crenças que são iniciadas e guiadas a partir de intenções, similarmente ao que ocorre com as ações intencionais. Tal desafio se deve a que, na teoria padrão da ação, para que uma ação corresponda a razões é preciso que ela seja iniciada e guiada por intenções.

A agência compartilhada é caracterizada de forma breve na medida em que ocorre quando dois ou mais indivíduos praticam uma ação em conjunto. Como exemplo podemos pensar em casos em que duas ou mais pessoas cantam uma canção ou carregam juntos algum objeto. A agência coletiva, por sua vez, acontece quando dois ou mais indivíduos agem como uma organização, “de acordo com certos princípios ou procedimentos que constituem e organizam o grupo” (SCHLOSSER, 2019, p. 19, tradução nossa).

Em se tratando tanto da agência coletiva quanto da agência compartilhada, Schlosser (2019) esclarece que é bastante questionado se tais tipos de agência podem ser reduzidos à agência dos indivíduos envolvidos ou se estão, de alguma forma, acima da agência individual. Nos termos da teoria padrão da ação, seria questionado se faz sentido atribuir estados e eventos mentais, como desejos, crenças e intenções, a grupos de indivíduos.

A agência relacional, por sua vez, foi postulada como uma tentativa inspirada em teorias feministas de reabilitar a autonomia como um valor, uma vez que, segundo críticas feministas, considerações tradicionais da autonomia são excessivamente individualistas e abstratas, pois não dão importância às relações interpessoais no desenvolvimento de um indivíduo autônomo. De acordo com Jane Dryden (2020, tradução nossa):

> Em geral, nas explicações da autonomia relacional, a autonomia é vista como um ideal pelo qual podemos medir quão bem um agente é capaz de negociar sua busca por objetivos e compromissos, alguns dos quais podem ser auto-escolhidos e outros como resultado de influências sociais e relacionais. Os laços sociais e relacionais são examinados em termos de seus efeitos sobre a competência de um agente nessa negociação: alguns fortalecem, outros criam obstáculos e outros são ambíguos. O foco principal da maioria das explicações de autonomia relacional, no entanto, tende a dirigir-se menos aos procedimentos e mais à mudança do modelo de eu autônomo, de um modelo individualista para outro que incorpora o contexto social.

Entendemos que a agência relacional diz respeito à agência autônoma. Assim, o agente se encontra inserido em um contexto social que deve ser levado em consideração na medida em que o contexto moldará e dará (ou não) condições para que o agente alcance/construa sua autonomia. Isso significa que o agente só será capaz de se desenvolver e de atingir seus objetivos na medida em que se relaciona com seu ambiente e com outros agentes.

Chegamos assim ao último item da lista: a agência artificial. De acordo com algumas considerações sobre a possibilidade de agência em sistemas artificiais, este tipo de agência não pode ser justificado, uma vez que sistemas artificiais não apresentam estados internos que seriam a base para os estados mentais representacionais responsáveis pela ação intencional. Porém, “se os sistemas artificiais não são capazes de agência intencional, conforme interpretado pela teoria padrão, eles ainda podem ser capazes de algum tipo mais básico de agência” (SCHLOSSER, 2019, p. 20, tradução nossa), segundo pressupostos da própria teoria.

Tendo em vista que temos um tópico reservado apenas para investigarmos a possibilidade de agência em sistemas artificiais, e que os outros tipos de agência foram elucidados de maneira satisfatória para nossos propósitos, damos continuidade à nossa investigação com a seção que tratará sobre a relação entre agência e autonomia.

Agência e autonomia

Embora a agência possa ser considerada em diversos contextos, a agência autônoma, característica considerada fundamental para que a agência possa ser pensada de maneira não trivial, possui características bastante específicas. Buscamos, assim, as bases que nos permitirão caracterizar o que é um agente

autônomo, e para tanto, faremos uso principalmente dos escritos de Sarah Buss e Andrea Westlund (2018).

No sentido comum do termo, um agente é aquele que pratica uma ação. Nesse sentido, praticar uma ação significa ter a possibilidade de iniciar um ato sem coerções externas. Ressaltam Buss e Westlund (2018, p. 2, tradução nossa) que apenas agentes têm essa capacidade e só será possível exercê-la caso o direito de ação exista:

> Uma vez que nada e nem ninguém tem o poder de agir, exceto o próprio agente, apenas ele tem o direito de exercer esse poder, se tiver o direito de agir. Isso significa que, na medida em que alguém é um agente, ou seja, na medida em que é uma pessoa que age - ela está correta em considerar seus próprios compromissos com o ato, seus próprios julgamentos e decisões sobre como deve agir enquanto autoridade.

Entendemos que para Buss e Westlund (2018), o poder de ação reside na parcela de autonomia necessária para que uma ação seja iniciada. E embora seja difícil encontrar concordância integral entre as diferentes teorias da agência, é possível perceber que estes estudos admitem algumas características essenciais da agência autônoma, tais como a possibilidade de justificar a própria ação e a capacidade de escolha, ou seja, a não coerção do agente para realizar a ação.

Buss e Westlund (2018) evidenciaram diversas características e problemáticas envolvidas na temática da agência autônoma. Para os nossos objetivos, destacaremos cinco dessas características como sendo fundamentais a todos os sistemas portadores de agência autônoma. São elas: (1) Todo agente tem uma autoridade sobre si mesmo que está fundamentada no fato de que o agente por si mesmo pode iniciar uma ação. (2) Para formar uma intenção de escolha, ainda que o agente deva seguir o comando ou o conselho de um outro, ele deve considerar seu próprio julgamento sobre como agir com autoridade. (3) O autogoverno mínimo parece não exigir nem mais nem menos do que ser o poder por trás de qualquer raciocínio que dê origem diretamente ao comportamento de alguém. (4) A coerência não é um fator necessário para a agência autônoma, ou seja, um agente não precisa sacrificar sua autonomia para decidir agir contrariamente aos seus compromissos e preocupações de longo prazo; agir “fora do personagem” não é condição suficiente para deixar de governar-se a si mesmo. Ainda que um agente considerado “fraco de vontade” não seja um exemplo paradigmático de alguém que se autogoverna quando

age, ele também desempenha um papel decisivo no poder relativo de seus próprios motivos; ele autoriza seu comportamento, embora acredite que tenha bons motivos para agir de outra forma.

> [...] basta observar que se a fraqueza da vontade é um fenômeno genuíno, então os agentes humanos têm a capacidade de se governar de uma maneira que eles próprios consideram injustificada. Eles podem afirmar uma autoridade sobre si mesmos que desafia a autoridade de sua própria razão (BUSS; WESTLUND, 2018, p. 15, tradução nossa).

Um caso de um agente fraco de vontade pode ser ilustrado com alguém que aceite ser corrompido; em casos de corrupção, os princípios individuais serão sempre questionados, porém não é incomum vermos que até os indivíduos que costumam se comportar da forma mais ética que podemos imaginar acabem por ceder a pressões exteriores em nome da própria índole ou imagem pessoal, em nome de grupos, e em nome do poder.

É possível afirmar que não é incomum observarmos agentes fracos de vontade, assim como não é incomum notar e reconhecer que cedemos aos nossos próprios princípios e vontades em nome daquilo que consideramos importante. Neste sentido, fazer parte de um grupo, manter as aparências, a manutenção de poderes e influências, proteger quem amamos e a nós mesmos, são casos comuns e cotidianos que podem fazer com o que os agentes mais íntegros ajam de forma contrária aos próprios princípios (talvez exemplificando situações de agência coletiva, no sentido proposto por Schlosser, 2019).

Por fim, a última característica destacada por Buss e Westlund (2018, p. 16) consiste (5) na capacidade de fazer planos. Segundo as autoras, os planos geralmente permitem que uma pessoa exerça alguma medida de controle sobre sua vida como um todo. Com efeito, uma pessoa pode se autogovernar em um determinado momento mesmo se desafiar suas tentativas anteriores de colocar restrições sobre como ela se governará nesse momento. A pessoa pode decidir abandonar seus planos ou modificá-los de maneiras que não havia previsto, e pode ainda rejeitar o conselho dos valores de longo prazo que fornecem a base lógica para esses planos.

Como foi dito, ainda existem muitas questões e controvérsias acerca da temática da agência, porém entendemos que as características apresentadas se

mostram suficientes para a especificação de um agente autônomo natural. Sendo elas, em síntese, a possibilidade de autorregulação sem coerções para planificar o futuro e/ou modificar tal planejamento.

Concluímos nossa investigação acerca da noção de agência autônoma em organismos naturais com as palavras de Buss e Westlund, que apontam que "[...] a única atitude da qual parece que nenhum agente pode ser alienado é o desejo de ter poder suficiente para determinar os próprios motivos - o desejo de ser um agente autônomo" (2018, p. 17, tradução nossa).

Mas será que sistemas artificiais poderiam ser considerados agentes nesse sentido forte, o qual inclui o desejo por autonomia?

Agência e autonomia em sistemas artificiais

No que diz respeito aos sistemas autônomos artificiais, a Declaração sobre Inteligência Artificial, Robótica e Sistemas 'Autônomos' (*Statement on Artificial Intelligence, Robotics, and 'Autonomous' Systems, European Group on Ethics in Science and New Technologies,* European Commission, 2018) estabelece que sistemas autônomos são sistemas que têm a capacidade de aprender sem o direcionamento ou o supervisionamento de seres humanos. A declaração em questão (2018) trata sobre tecnologias digitais autônomas, inteligência artificial e aprendizado de máquinas que são capazes de redefinir ou aprimorar as condições de trabalho humano para além de sua programação inicial.

A *European Commission* (2018) enfatizou que é uma infelicidade o fato de algumas das ferramentas cognitivas mais poderosas em questão permanecerem ainda obscuras, uma vez que suas ações não são mais programadas por humanos em um sentido linear. Por exemplo:

> O *Google Brain* desenvolve IA que supostamente constrói IA melhor e mais rápido que os seres humanos. O *AlphaZero* pode se auto inicializar em quatro horas, ignorando completamente as regras do xadrez e o nível do campeão mundial. É impossível entender exatamente como *AlphaGo* conseguiu vencer o campeão humano de *Go World*. O aprendizado profundo e as chamadas "abordagens geradoras de redes contraditórias" permitem que as máquinas "ensinem" a si mesmas novas estratégias e busquem novas evidências para analisar. Nesse sentido, suas ações geralmente não são mais inteligíveis e não estão mais abertas ao escrutínio dos seres humanos. Esse é o caso porque, primeiro, é impossível estabelecer como eles alcançam seus resultados além dos algoritmos iniciais. Segundo, seu desempenho é

> baseado nos dados que foram usados durante o processo de aprendizagem e que podem não estar mais disponíveis ou acessíveis. Assim, preconceitos e erros apresentados no passado ficam enraizados no sistema (E.C., 2018, p. 6, tradução nossa).

Entendemos as preocupações dos autores do documento em questão ao afirmarem seu desagrado com o fato dessas ferramentas se comportarem obscuramente, porque, uma vez que sistemas artificiais inteligentes têm a capacidade de se programarem e de programar outros sistemas, os agentes naturais tendem a perder sua capacidade de controle sobre eles. Embora a sociedade esteja sempre em busca de desenvolvimento e aprimoramento, e principalmente de progresso tecnológico, quando se trata de ferramentas que em algum grau passam a ser independentes de seus criadores, corre-se o perigo de se tornar refém de suas próprias criações.

Em síntese, sistemas capazes de aprender tarefas sem a supervisão de um ser humano, ou seja, sem que sejam programados especificamente para realizar tarefas determinadas, podem ser considerados autônomos. Eles podem se manifestar como sistemas robóticos altamente tecnológicos ou como *softwares* inteligentes. “Muitos deles são liberados para o mundo sem supervisão e podem realizar coisas que não são previstas por seus projetistas ou proprietários humanos” (E.C., 2018, p. 7, tradução nossa).

A declaração europeia sobre sistemas inteligentes enfatiza ainda que a autonomia está relacionada diretamente à dignidade humana e à capacidade humana de autorregulação, englobando também o direito de ser livre para escolher seus padrões, objetivos e propósitos.

> Os processos cognitivos que apoiam e facilitam essas capacidades estão entre os mais intimamente identificados com a dignidade das pessoas humanas, agência e atividade humana por excelência. Eles normalmente envolvem as características da autoconsciência e da auto autoria de acordo com razões e valores (E.C., 2018, p. 9, tradução nossa).

Assim sendo, para a Comissão Europeia (2018), ainda que o conceito de autonomia tenha ganhado espaço nos debates científicos e no debate público para fazer referência a elevados graus de automação e de independência dos seres humanos em relação à operabilidade e à tomada de decisões, o termo autonomia não

poderia ser empregado em outros sentidos e deveria ser utilizado apenas quando se trata de seres humanos.

A declaração afirma também que a posição moral e a dignidade humana não devem ser transferidas aos sistemas artificiais. Assim, os seres humanos deveriam ser sempre responsáveis por suas criações e permanecerem no controle das tecnologias.

> Os seres humanos devem ser capazes de determinar quais valores são servidos pela tecnologia, o que é moralmente relevante, quais objetivos finais e concepções do bem são dignos de serem buscados. Isso não pode ser deixado para máquinas, não importa o quão poderosas elas sejam (E.C., 2018, p. 10, tradução nossa).

Partindo do fato de que a postura adotada pela Comissão Europeia é bastante antropocêntrica em suas considerações a respeito de autonomia, nos questionamos até que ponto ela é válida quando estamos tratando de sistemas artificiais cujos *outputs* parecem ir além de sua programação inicial, como resultado do desenvolvimento de capacidades de ajuste desses *outputs* por meio de aprendizado. Tais sistemas artificiais parecem ser capazes de produzir resultados e respostas que não foram previstas pelos seus projetistas e programadores.

Consideramos muito problemático simplesmente afirmar que é possível e aceitável transferir a responsabilidade de uma pessoa para uma máquina ou para um *software*, uma vez que entendemos que pessoas podem, de fato, ser culpabilizadas e responsabilizadas por suas ações e decisões, na medida em que têm a capacidade de reflexão sobre as mesmas. Porém quando levamos em consideração os avanços tecnológicos e em como vidas humanas são colocadas em poder da tecnologia, que é projetada para ser cada vez mais autônoma e inteligente, é necessário refletir cuidadosamente como as decisões de um sistema artificial podem afetar vidas humanas.

Enfatizamos que, na medida em que um sistema artificial é capaz de tomar decisões e modificar o rumo de sua programação inicial sem a necessidade da intervenção de um supervisor humano, esse sistema poderia sim ser considerado autônomo, ainda que em um sentido fraco, e não necessariamente no sentido de realizar um tipo de reflexão de segunda ordem semelhante ao “ser capaz de refletir sobre as próprias decisões” dos agentes humanos.

A fim de complementar a nossa reflexão, apresentamos algumas considerações de Franklin e Graesser (1996), proponentes de definições e esclarecimentos sobre o que viria a ser a agência autônoma que vai ao encontro da definição proposta pela declaração europeia[2].

Para os autores existem dois tipos de usos comuns para a palavra agente: o primeiro se refere a aquele que age, ou que pode agir, e o segundo uso da palavra se refere àquele que age, após permissão, no lugar de outro agente. “Como ‘aquele que age no lugar de’ age, o segundo uso requer o primeiro” (FRAKLIN; GRAESSER, 1996, p. 25, tradução nossa).

Assim, para ilustrar o primeiro uso do termo “agência” e dar exemplos que podem sustentar a definição que está sendo construída, os autores apresentam três tipos de agentes: os seres humanos e os animais, os *softwares agents* que “vivem”[3] em sistemas operacionais de computadores, banco de dados e redes, e por fim agentes artificiais (*artificial life agents*) que “vivem” em ambiente artificiais, na tela de um computador ou em sua memória. E os requisitos que constituem a “essência” de um agente autônomo são:

> Cada um está situado e faz parte de algum ambiente. Cada um percebe o seu ambiente e atua autonomamente sobre ele. Nenhuma outra entidade é necessária para alimentar seus *inputs*, ou para interpretar e usar seus *outputs*. Cada um age em busca de sua própria agenda, seja satisfazendo os impulsos com que evoluíram, como em seres humanos e animais, ou perseguindo metas criadas por algum outro agente, como em agentes de *software*. (Agentes de vida artificial podem ser de qualquer uma dessas variedades). Cada um age de tal modo que suas ações atuais podem afetar sua percepção posterior, isto é, suas ações afetam seu ambiente. Finalmente, cada um age continuamente ao longo de algum período de tempo (FRANKLIN; GRAESSER, 1996, p. 25, tradução nossa).

Entendemos que para Franklin e Graesser (1996), o que os agentes autônomos compartilham como característica comum é a necessidade de estarem situados e fazerem parte de um ambiente na medida em que cada agente pode perceber este ambiente e agir de maneira autônoma sobre ele, seguindo uma agenda própria e podendo modificar o seu futuro.

---

[2] A declaração da Comissão Europeia (2018) trata de *softwares*, programas e sistemas inteligentes atualmente elaborados ou já tecnicamente concebíveis, enquanto o trabalho de Franklin e Graesser, publicado em 1996, contribui com a questão conceitual envolvida na problemática que envolve a agência autônoma de sistemas artificiais.

[3] Grifo dos autores.

A fim de esclarecimento, consideramos importante pontuar ainda a diferença entre simples programas de computadores e os agentes artificiais em si, uma vez que ao utilizarmos nossos dispositivos inteligentes para navegação e resolução de tarefas, não sabemos realmente com qual tipo de situação estamos lidando.

Franklin e Graesser (1996, p. 26) afirmam que um programa simples passa a funcionar quando ativado por um usuário: após cumprir a tarefa, o usuário do programa encerra suas atividades até que seja ativado novamente. Os autores usam como exemplo um programa de folha de pagamentos: esses programas não são autônomos, uma vez que seus *outputs* não alteram ou não agregam informação para que se comportem de maneira diferente do que foi inicialmente previsto.

Por fim, é possível afirmar resumidamente que para um sistema ser considerado um agente ele deve estar situado em um ambiente e ser dependente das condições que determinado ambiente fornece para que ele possa atuar nesse meio. Ademais, um agente pode ser considerado autônomo quando é capaz de escolher suas ações de forma independente enquanto cumpre sua agenda: “Agentes autônomos estão situados em algum ambiente. Altere o ambiente e talvez não tenhamos mais um agente. Um robô com sensores apenas visuais em um ambiente sem luz não é um agente” (FRANKLIN; GRAESSER, 1996, p. 26, tradução nossa).

Entendemos, assim, que agentes artificiais não devem ser caracterizados como agentes autônomos no sentido forte do termo, ou seja, agentes conscientes de sua capacidade de ação e de escolha. Porém, se um agente artificial, ainda que a cumprir ordens, exiba comportamentos e *outputs* que não foram previstos, se ele é capaz de escolher entre opções de acordo com o ambiente e contexto no qual está inserido, indo além de sua programação inicial e aprendendo com seu ambiente, esse agente pode ser considerado como agente portador de algum grau de autonomia.

Considerações Finais

Tecemos, ao longo do texto, uma investigação acerca do conceito de agência com o objetivo de verificar se sistemas artificiais podem ser incluídos na categoria de agentes e de que modo tal consideração pode ser realizada de maneira criteriosa, a fim de evitar leviandade.

Verificamos que para Anscombe (1957) e Davidson (1963), a noção de agência se sustenta a partir da noção de ação intencional. Neste sentido, apenas organismos com capacidades cognitivas de segunda ordem seriam capazes de ser considerados agentes no sentido rigoroso do conceito.

Porém, foi possível conferir também que existem diferentes tipos de agência que não requerem a atribuição de representações mentais (SCHLOSSER, 2019). Neste sentido, entidades que não possuem capacidades cognitivas ditas de segunda ordem podem também possuir algum tipo legítimo de agência.

Assim, após observar noções tanto padrão quanto alternativas de agência, e de caracterizar noções de ação e de autonomia, foi possível verificar que sistemas artificiais, em determinados contextos, cumprem com as condições necessárias para serem considerados como sistemas portadores tanto de agência quanto de (algum grau de) autonomia. E ainda que muitas das características de um agente natural autônomo não se apliquem a um agente artificial autônomo, tal fato não deve ser tomado como condição para excluir os agentes artificiais da categoria de agentes autônomos.

Para concluir, corroboramos a posição da declaração europeia que afirma que a posição moral e a dignidade humana não devem ser transferidas a sistemas artificiais, não importando o quanto estes sejam considerados inteligentes, de modo que os seres humanos devem sempre se manter responsáveis e no controle daquilo que criam.

Referências

ANSCOMBE, E. G. M. **Intention**. 2. ed. England: Harvard University Press, 1957.

BUSS, S.; WESTLUND, A. Personal Autonomy. *In:* ZALTA, E. N. **The Stanford Encyclopedia of Philosophy.** Spring 2018 Edition. Disponível em: https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/.

DAVIDSON, D. Actions, Reasons, and Causes. *In*: **Essays on Actions and Events**. 2. ed. Oxford: Clarendon Press, 2002. p. 3-19.

DAVIDSON, D. Agency. *In:* **Essays on Actions and Events**. 2. ed. Oxford: Clarendon Press, 2002. p. 43-61.

DRYDEN, J. Autonomy. *In:* **Internet Encyclopedia of Philosophy**. Disponível em: http://iep.utm.edu/autonomy/. Acesso em: 1.abr.2024.

EUROPEAN COMMISSION. **Artificial Intelligence, Robotics and 'Autonomous' Systems.** Luxembourg: Publications Office of the European Union, 2018. Disponível em: https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf. Acesso em: 15.out.2018.

FRANKLIN, S.; GRAESSER, A. *Is it an Agent, or just a Program?* A Taxonomy for Autonomous Agents. Institute For Intelligent Systems - University of Memphis. **Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages.** Springer-Verlag, 1996.

HALDANE, J. "*In Memoriam: G. E. M. Anscombe (1919-2001)*". ***The Review of Metaphysics*, v. 53**, n. 4, p.1019–1021, 2000. Disponível em: https://www.st-andrews.ac.uk/~jjh1/Haldob.pdf.

SCHLOSSER, M. Agency. *In:* ZALTA, E. N. (Ed.). **The Stanford Encyclopedia of Philosophy**. Winter 2019 Edition. Disponível em: https://plato.stanford.edu/archives/win2019/entries/agency/.

SPEAKS, J. Davidson's "Actions, Reasons, and Causes". **University of Notre Dame**, 2009.

WEISER, M. The computer for the 21st century. **Scientific American**, v. 265, n. 3, 1991, p. 94-105.

WISEMAN, R. *Intentional action* in Anscombe's Intention. **Routledge philosophy guidebook to Anscombe's Intention**, n. 15, 2016, p. 77-111.

Recebido: 29/04/2024
Aprovado: 29/06/2024