Revista de Estudantes de Filosofia da Universidade de Brasília

10.26512/pl.v13i30.55527

A AGÊNCIA MORAL ENTRE A INTELIGÊNCIA ARTICIFIAL E ROBÔS

MORAL AGENCY BETWEEN ARTIFICIAL INTELLIGENCE AND ROBOTS

Maurício Cavalcante Rios[1]
Instituto Federal da Bahia

[1]Doutor em Ensino, Filosofia e História das Ciências pela Universidade Federal da Bahia (UFBA). Professor de Filosofia do IFBA.
E-mail: mauricioriosfil@gmail.com.
Lattes: http://lattes.cnpq.br/7639267588949243. Orcid: https://orcid.org/0000-0002-1652-9318.

RESUMO: Este artigo tem como objetivo compreender filosoficamente as relações entre inteligência artificial, robôs e agência moral. O avanço da inteligência artificial tem modificado as formas de pensamento e comportamento dos seres humanos em sociedade. Para o nosso trabalho, partiremos do seguinte problema: se atribuímos à inteligência artificial uma agência moral, então a inteligência artificial pode ser moralmente responsável? Nossa hipótese a ser verificada é que se a inteligência artificial simula ações morais na tomada de decisões, então uma agência moral apenas simula formas de responsabilidade moral. Esta proposta justifica-se devido à importância contemporânea da filosofia para a compreensão e o debate de conceitos teóricos que envolvem a inteligência artificial. Esses conceitos teóricos são os agentes inteligentes, os algoritmos, o machine learning (aprendizado de máquina), o *big data* (grande quantidade de dados computáveis), a cognição, a informação, a lógica, as redes neurais, dentre outros.

Palavras-chaves: Inteligência Artificial. Agentes Inteligentes. Agência Moral. Robôs. Responsabilidade

ABSTRACT:

This article aims to understand the relationships between artificial intelligence, robots and moral agency philosophically. The advance of artificial intelligence has changed the forms of thought and behavior of human beings in society. For our work, we will start from the following problem: if we attribute to artificial intelligence a moral agency, then artificial intelligence can be morally responsible? Our hypothesis to verify is that if artificial intelligence simulates moral actions in decision making, then a moral agency only simulates forms of moral responsibility. This proposal is justified due to the contemporary importance of philosophy for the understanding and debate of theoretical concepts involving artificial intelligence. These theoretical concepts are intelligent agents, algorithms, machine learning (machine learning), big data (large amount of computable data), cognition, information, logic, neural networks, among others.

Keywords: Artificial Intelligence. Smart Agents. Moral Agency. Robots. Responsibility.

INTRODUÇÃO

O objetivo deste artigo é compreender filosoficamente as relações entre inteligência artificial, robôs e agência moral. O avanço da inteligência artificial tem modificado as formas de pensamento e comportamento dos seres humanos em sociedade. Não apenas estamos falando de uma sociedade informacional, sociedade em rede ou sociedade cibernética, mas de uma sociedade que tornou possível a existência de sistemas inteligentes ou de sistemas que simulam a inteligência humana.

Mediante a isso, é necessário refletir sobre o que seja a inteligência artificial (Teixeira, 2014), abordando sua filosofia, paradigmas e uma questão central: máquinas podem pensar? Este problema foi lançado por Alan Turing (1950) e desenvolvido por uma diversidade de perspectivas que concordam ou não com a capacidade das máquinas pensarem (Bringsjord, 2018).

Nosso trabalho parte do seguinte problema: se atribuímos à inteligência artificial uma agência moral, então a inteligência artificial pode ser moralmente responsável? Nossa hipótese a ser verificada é que se a inteligência artificial simula ações morais na tomada de decisões, então uma agência moral apenas simula formas de responsabilidade moral. Para compreender essa hipótese, é necessário entender o que é a inteligência artificial, a agência inteligente artificial, robôs e responsabilidade. Destacamos que não há uma linearidade de desenvolvimento filosófico, científico e técnico na história da inteligência artificial. Pretendemos, ao afirmar isso, que a história da inteligência artificial é repleta de paradigmas, rupturas e descontinuidades teóricas e epistemológicas.

Na primeira seção, utilizaremos os conceitos de Stuart Russel e Peter Novig (2010) sobre Inteligência Artificial e Marc Champagne (2021) sobre agência inteligente para o entendimento de conceitos como: a) Racionalidade ideal; b) Agentes inteligentes e c) Responsabilidade moral dos agentes inteligentes. Na segunda seção, trataremos de analisar a agência moral artificial a partir da visão padrão e funcionalista. Para isso, utilizaremos: a) Os conceitos de Angeluci e Rossetti (2021) sobre a relação entre ética, algoritmos e aprendizado de máquina; c) As análises de Bringsjord (2018) e Mcnamara (2021) sobre deontologia aplicada à inteligência artificial e d) As definições de Behdadi e Munthe (2021) sobre a visão padrão e funcionalista da inteligência artificial. Na terceira seção, abordaremos os argumentos de Muller (2020) sobre a responsabilidade dos robôs.

Nas considerações finais, consideramos que as diferentes conceituações sobre responsabilidade moral, agência moral e consciência nos leva a um posicionamento atribuicionista. Reconhecemos que o problema metafísico-ontológico entre seres humanos e entidades inteligentes artificiais ainda não está resolvido.

1 AGENTES INTELIGENTES ARTIFICIAIS E RESPONSABILIDADE

A reflexão sobre a inteligência artificial tornou-se relevante para a filosofia contemporânea, uma vez que há impactos cruciais sobre novas concepções de pensamento (máquinas podem pensar?), epistemológicas (agentes artificiais inteligentes geram conhecimento e aprendem?), éticas (agentes artificiais inteligentes têm responsabilidade moral?), lógicas (a inteligência artificial necessita de novas lógicas aplicadas) e ontológicas (até que ponto, poderemos distinguir humanos de entidades artificiais inteligentes?).

Os agentes inteligentes, os algoritmos, o *machine learning* (aprendizado de máquina), o *big data* (grande quantidade de dados computáveis), a cognição, a informação e as redes neurais constituem elementos fundamentais para a compreensão teórica e funcional da inteligência artificial (Bringsjord, 2018; Floridi, 2011; O'Neil, 2017, 2020). A filosofia dirige-se a esses elementos fundamentais como conceitos teóricos, procurando não só elucidá-los, mas investigando quais tipos de relações podem ser feitas entre esses conceitos e o pensamento, o corpo e os artefatos humanos. Para isso, diversos campos da filosofia buscam compreender essas relações: a) A epistemologia computacional (Dantas, 2019); b) A filosofia da informação (Floridi, 2011); c) As teorias da mente (Tripicchio, 2003); d) A filosofia da ciência (Korb, 2004), e) A filosofia da tecnologia (Cupani, 2016) e f) A filosofia da inteligência artificial (Bringsjord, 2018). Dentre essas perspectivas sobre a inteligência artificial, encontramos as seguintes: a) Representacionistas: a inteligência artificial é uma representação de modelos mentais; b) Conexionistas: a inteligência artificial corresponderia aos fenômenos neurais e c) Fenomênicas: a inteligência artificial passa a ser um objeto ou processos fenomênicos.

A diversidade de estudos filosóficos sobre a inteligência artificial leva a questões para os debates científicos como, por exemplo, nas ciências da computação, nas engenharias, nas neurociências e em outras especialidades que trabalham com inteligência artificial. Uma dessas questões é a seguinte: máquinas podem pensar? (Turing, 1950).

Destacamos que um dos papéis da filosofia é compreender o que é pensar e, nesse sentido, é de profundo interesse filosófico, pois o entendimento do significado de pensamento ou de formas de pensar é útil para o desenvolvimento da inteligência artificial:

> Russell e Norvig [...] fornecem um conjunto de respostas possíveis para "O que é IA?" pergunta que tem progresso no próprio campo. Todas as respostas assumem que a IA deve ser definida nos termos de seus objetivos: uma definição candidata tem a forma "IA é o campo que visa construir ..." As respostas se enquadram em um quarteto de tipos colocados ao longo de duas dimensões. Uma dimensão é quando o objetivo é corresponder ao desempenho humano ou, em vez disso, racionalidade

> ideal. A outra dimensão é quando o objetivo é construir sistemas que raciocinam/pensam, ou melhor, sistemas que agem (Bringsjord, 2018, n.p.).

O sentido das respostas nos leva a entender que a inteligência artificial é um campo que visa a construção de agentes que simulam a inteligência humana. Dessa forma, esses agentes atuam sobre um meio procurando seguir uma racionalidade ideal. As ações e decisões tomadas pelos agentes inteligentes partem de padrões estabelecidos por um modelo que segue uma racionalidade. O quarteto que define as dimensões da inteligência artificial é representado abaixo da seguinte forma:

Tabela 1 – *Dimensões da Inteligência Artificial*

	Baseado em Humano	**Racionalidade Ideal**
Baseada em Raciocínio	Sistemas que pensam como humanos	Sistemas que pensam racionalmente.
Baseada em Comportamento	Sistemas que agem como humanos.	Sistemas que agem racionalmente.

Fonte: AIMA (2018)[2]

Notemos que há inteligências artificiais baseadas em raciocínio e baseadas em comportamento. Isso gera duas formas de cognição para essas inteligências: agir ou raciocinar. Essas dimensões ou simulam o ser humano ou seguem um modelo ideal de racionalidade. Na dimensão da inteligência artificial baseada em raciocínio, compreendemos os sistemas que simulam a inteligência humana ou sistemas que podem raciocinar por conta própria. Ambas as formas executam formalismos e operações lógicas. Nesta dimensão, o paradigma central é o modelo mental racional. Na dimensão da inteligência artificial baseada em comportamento, entendemos que os sistemas executam ações por simulação do comportamento humano ou sistemas que executam ações por conta própria.

A partir do quarteto acima destacado, observamos que o modelo que simula o comportamento humano pode ser interpretado como uma inteligência artificial fraca: "A IA "fraca", por outro lado, procura criar máquinas de processamento de informações que parecem ter o repertório mental completo de pessoas humanas" (Searle, 1997, pp. 09-10). Por sua vez, o modelo que segue uma racionalidade ideal pode ser interpretado como uma inteligência artificial forte: "Máquinas que têm todos os poderes mentais que temos, incluindo a consciência fenomenal" (Bringsjord, 2018, n.p.). Há uma série de questões filosóficas sobre essas formas de inteligência artificial, uma vez que não

[2] Esse quadro é apresentado por Bringsjord (2018) em seu artigo – *Artificial Intelligency* – onde sintetiza as ideias de Russel e Novig (2010) na obra AIMA (*Artificial Intelligence: A Modern Approach*). *Cf.* https://plato.stanford.edu/entries/artificial-intelligence/#WhatExacAI.

conhecemos profundamente os mecanismos de manifestação da consciência ou se é possível atribuir um *self* ou uma entidade que reflita sobre si mesmo.

Além de uma consciência, o núcleo da inteligência artificial forte é um modelo de racionalidade ideal que executa operações lógicas e age sob a tomada de decisões por si mesma (agentes racionais). Mas o que vem a ser essa racionalidade ideal quando relacionada à inteligência? Ela envolve percepção do meio? Vejamos:

> E quanto a Russell e Norvig? Qual é a resposta deles para o Que é IA? [...] Eles são firmes na colocação de "atuar racionalmente". De fato, é seguro dizer que eles são os principais defensores desta resposta e que foram apologistas notavelmente bem-sucedidos. Sua série AIMA extremamente influente pode ser vista como uma defesa e especificação de livros da categoria Ideal/Agir. Observaremos, um pouco mais tarde, como Russell e Norvig estabelecem toda a IA em termos de agentes inteligentes, que são sistemas que agem de acordo com vários padrões ideais de racionalidade. Mas primeiro vamos olhar um pouco mais perto a visão sobre inteligência subjacente ao texto do AIMA. Podemos fazê-lo recorrendo a Russell [...]. Aqui, Russell reformula a questão "O que é IA?" como a questão "O que é inteligência?" (presumivelmente sob a suposição de que temos uma boa compreensão do que é um artefato) e, em seguida, ele identifica inteligência com racionalidade. Mais especificamente, Russell vê a IA como o campo dedicado à construção de agentes inteligentes, que são funções que tomam como tuplas[3] de percepções do ambiente externo e produzindo comportamento (ações) com base nessas percepções (Bringsjord, 2018, n.p.).

Na inteligência artificial, a agência inteligente relaciona-se com a construção de artefatos que atuam racionalmente sobre um meio a partir de *inputs* e *outputs*. Vemos que não se trata apenas de seguir padrões de programação complexos, mas se trata da construção de agentes inteligentes que percebem o meio e atuam sobre ele. O modelo básico de inteligência/racionalidade consiste na ação de um agente sobre um meio que produz uma sequência de estados (memórias) sobre esse meio. Essa sequência é avaliada por determinadas medidas a fim de calcular sua utilidade. Esse modelo básico não envolve apenas operações lógicas, mas a entrada e saída de dados provenientes de estados percebidos do meio. Esse modelo não deixa de ser um teste de inteligência, uma vez que Russel e Novig (2010, p. 02) apresentam um esquema mais concreto dos objetivos da inteligência artificial em sua história: "A IA é o campo dedicado à construção de artefatos inteligentes, onde 'inteligente' é operacionalizado por meio de testes de inteligência (como a Escala de Inteligência de Adultos de Wechsler) e outros testes de capacidade mental (incluindo, por exemplo, testes de capacidade mecânica e criatividade)" (Newell, 1973 *apud* Bringsjord, 2018, n.p.).

Ao percebermos que os testes de inteligência são aplicados para o desenvolvimento da inteligência artificial, podemos declarar que há uma psicometria? No sentido de que os modelos de

[3] Tuplas são sequências de dados imutáveis e heterogêneos.

inteligência artificial são baseados em nossos modelos mentais, os testes de inteligência utilizam técnicas quantitativas para avaliar dados de performance na solução de problemas. Da mesma forma, isso pode ser aplicado à inteligência artificial a fim de avaliar seu desempenho no conjunto de suas ações, funções e decisões. A inteligência artificial psicométrica não apenas se baseia nos restritos testes de quociente de inteligência, mas envolve outras áreas:

> A IA psicométrica é o campo dedicado à construção de entidades de processamento de informações capazes de pelo menos um desempenho sólido em todos os testes estabelecidos e validados de inteligência e capacidade mental, uma classe de testes que inclui não apenas os testes de QI bastante restritivos, mas também considerando que isso é importante, dados os testes nos quais os trabalhos da presente edição especial[4] são focados - testes de criatividade artística e literária, capacidade mecânica e assim por diante (Bringsjord, 2018, n.p.).

Habilidades artísticas, criativas e mecânicas também devem ser estudadas pelo campo da inteligência artificial. Para isso, faz-se necessário ampliar os testes para outras capacidades cognitivas além dos testes que avaliam a capacidade abstrata, formal, geométrica, lógica e matemática, mas isso não implica em declarar que sistemas de inteligência artificial criam ou possuem dotes artísticos e literários. Percebemos que os dados, o processamento dos dados, a organização de informações e a ação de algoritmos ainda estão baseados em aspectos lógicos. Embora possamos considerar a existência de uma inteligência artificial não-logicista, isso não é a negação de todo e qualquer formalismo. A inteligência artificial não-logicista relaciona-se com formalismos específicos e não com o sistema formal como um todo. Nessa categoria de inteligência artificial, incluímos as abordagens probabilísticas e neurocomputacionais. A abordagem probabilística baseia-se em sistemas que calculam hipóteses para resolver problemas em um largo espectro de variáveis. A inteligência artificial baseada em redes neurais baseia-se em arquitetura de sistemas voltados para a aprendizagem e são constituídos por unidades que se assemelham a neurônios (Bringsjord, 2018). As conexões entre essas unidades são chamadas de *links* com funções de entrada e saída de dados em relação ao meio. Nossas perguntas são: como essa arquitetura de inteligência artificial pode tomar decisões responsáveis? Isso é possível? Agentes inteligentes artificiais são responsáveis?

A atribuição de responsabilidade moral aos agentes inteligentes deve passar por um crivo metafísico e ontológico: até que ponto, podemos diferenciar seres humanos de sistemas de inteligência artificial:

[4] Trata-se do artigo - *Psychometric artificial intelligence* – publicado no *Journal of Experimental & Theoretical Artificial Intelligence.*

> A distinção real/artificial é real, não artificial. Afinal, não é preciso atribuir uma "alma" ou "mente" para o ouro para insistir em sua diferença no ouro de tolo. A ênfase em culpar ou elogiar somente agentes que são realmente responsáveis se encaixa nas práticas comuns. Podemos errar em determinar o que exibe uma agência real – creditando erradamente a um albatroz uma boa navegação [...], mas depois de descobrirmos o erro de nossos caminhos, paramos de manter o alvo na questão responsável. Para tornar isso concreto, aqueles de nós que pensam que Deus existe não temem sua ira ou se preocupam com o curso de ação que ele supostamente prescreve. Para onde vai a ontologia, a normatividade normalmente segue.
> Criamos robôs, então sabemos que eles não estão no comando (ou pelo menos refletem nossas escolhas - incluindo a escolha de construí-los em primeiro lugar). Certamente, somos livres para cunhar novos conceitos para se encaixar em nossas novas criações, mas se tais conceitos terão algum valor na imaginação humana, isso é outro assunto. Portanto, suponho que dilemas emergentes sobre como tratar os robôs forçarão a distinção verdadeiro/artificial assumir seu lugar ao lado dos outros pilares identificados pela psicologia moral, como cuidados/danos, justiça/trapaça, lealdade/traição, autoridade/subversão e santidade/degradação. O que resta para ser visto é se a artificialidade é uma consideração para ponderar entre outras, ou se sua presença torna todos os outros padrões morais inaplicáveis. Afinal, não lamentamos realmente a morte de pessoas que aparecem nos filmes porque sabemos muito bem que, por mais convincente que seja a atuação, aqueles personagens são falsos. Somos, por outro lado, profundamente movidos por uma foto de um garoto sírio morto de bruços em uma praia porque é um exemplo real (Champagne, 2021, p. 81).

A distinção ontológica entre o real e o artificial consiste na ideia de que os artefatos são objetos produzidos pelos seres humanos, mas que possuem uma forma e conteúdo material e virtual diferente do ser humano. Embora possamos declarar que os artefatos são elaborados de materiais provenientes da natureza, eles possuem um *design*[5] e uma função. O processo de *design* é uma prática que envolve a fabricação de um artefato desde a escolha de seus materiais, requisitos funcionais, prototipagem até a reprodução e correções tecnológicas. O artefato é resultado de um processo que está sujeito às alterações de funcionamento de acordo com a suas melhorias ao meio e público. Um artefato é criado para ser um produto ou componente de outro artefato maior. Nesse aspecto, os artefatos possuem funções, pois servem para algo e a mente humana aplica intenções a essas funções. Percebemos que há uma dualidade entre uma mente humana dotada de intencionalidade e um artefato que possui uma função. Ainda que a noção de função seja central para os artefatos, pode-se ainda declarar que esta noção está presente na biologia (determinadas organelas citoplasmáticas exercem funções), mas sem intencionalidade humana. No caso dos artefatos, a intencionalidade humana está presente, uma vez que dependem de nossas ações responsivas. Mas como fica o caso dos agentes inteligentes artificiais? Em termos de uma filosofia da tecnologia, esses agentes seguem um processo de *design*, mas até que ponto, uma função de um artefato inteligente é ligada à intencionalidade humana?

[5] Para maiores detalhes, conferir o problema do design em: https://plato.stanford.edu/archives/fall2024/entries/technology/.

Entendemos que atribuímos aos agentes inteligentes artificiais um senso de responsabilidade, mas isso significa que o agente é moralmente responsável por suas ações? Consideramos isso porque os sistemas de inteligência artificial não possuem um *self* ou entidade autorreflexiva, uma consciência onde podem discernir valores morais para a tomada de decisões. Dizer que um agente inteligente artificial cometeu um crime, por exemplo, é atribuir uma consciência moral que não é real nesses artefatos tecnológicos. Podemos sim utilizá-los para nossas decisões, mas não é o agente inteligente artificial que age intencionalmente. A agência real mantém a distinção ontológica entre seres humanos e inteligência artificial, pois temos uma consciência responsiva.

2 AGÊNCIA MORAL ARTIFICIAL: VISÃO PADRÃO E FUNCIONALISTA

A agência moral artificial refere-se à capacidade de um sistema de inteligência artificial atuar como um agente moral, ou seja, tomar decisões éticas e comportar-se de maneira que reflita padrões morais e éticos estabelecidos. Este conceito envolve várias camadas de complexidade, incluindo considerações técnicas, filosóficas e sociais. A tomada de decisão pela inteligência artificial, por exemplo, avalia cenários a partir de princípios éticos definidos nos algoritmos:

> Algoritmos de tomada de decisão são usados em uma variedade de domínios, de modelos simplistas de tomada de decisão [...] a algoritmos de perfil complexos [...]. Exemplos contemporâneos notáveis incluem agentes de software online usados por provedores de serviços online para realizar operações em nome dos usuários [...]; algoritmos de resolução de disputas online que substituem os tomadores de decisão humanos na mediação de disputas [...]; sistemas de recomendação e filtragem que comparam e agrupam usuários para fornecer conteúdo personalizado [...]; sistemas de apoio à decisão clínica [...] que recomendam diagnósticos e tratamentos aos médicos [...]; e sistemas de policiamento preditivo que prevêem pontos críticos de atividades criminosas (Mittelstadt *et al.*, 2016 *apud* Angeluci e Rossetti, 2021, p. 12).

A tomada de decisão é um dos principais temas da inteligência artificial quando nos referimos às implicações éticas filosóficas e concretas. Como podemos observar, as aplicações da tomada de decisão estão presentes nas redes sociais, no entretenimento (*games*, *streaming*, realidade virtual, dentre outros), na medicina e na segurança pública. Nossa pergunta é como a inteligência artificial "aprende" para desenvolver agenciamentos morais e éticos? Devemos compreender a aprendizagem de máquina/algoritmos.

Um sistema de inteligência artificial pode treinar a execução de ações a partir de princípios éticos de forma contínua. Para que isso aconteça, os algoritmos devem ser organizados para funcionar em determinados padrões e regras, programando continuamente princípios éticos. Dizemos que os

algoritmos "aprendem" através da alimentação de grande quantidade de dados coletados. Isso representa um retorno para a aprendizagem dos algoritmos em cenários que envolvam dilemas morais. A aprendizagem de máquinas/algoritmos envolve autonomia do sistema, incertezas e desafios éticos:

> O aprendizado da máquina é definido como a capacidade de definir ou modificar regras de tomada de decisão de forma autônoma. As capacidades de aprendizado conferem aos algoritmos algum grau de autonomia. O impacto dessa autonomia deve permanecer incerto, mas até certo ponto. Como resultado, as tarefas executadas pelo aprendizado de máquina são difíceis de prever antecipadamente (como uma nova entrada será tratada) ou explicadas posteriormente (como uma determinada decisão foi tomada). A incerteza pode, assim, inibir a identificação e a correção de desafios éticos no projeto e operação de algoritmos de aprendizagem, que são capazes de tomar decisões. Então a questão fundamental que surge é: essa tomada de decisão algorítmica, por ser autônoma, é fundada em quais valores éticos? (Angeluci e Rossetti, 2021, p. 04).

A ética da inteligência artificial é um subcampo derivado da ética da computação e toma como referencial as ações humanas. A ética da inteligência artificial considera a autonomia dos agentes inteligentes artificiais em tomar decisões. Para isso, deve existir uma estrutura em que se possa construir agentes inteligentes artificiais a partir de um código moral (Bringsjord, 2018). Essa estrutura deve ter características formais partindo de algo similar aos nossos raciocínios. O campo que estuda a aplicação dessa estrutura formal é a lógica deôntica cuja característica parte de um código moral. Como exemplo, podemos citar, no quadro abaixo, palavras que sintetizam comandos deontológicos:

Tabela 2 – *Lógica Deôntica: noções*

a) Permissível	b) Inadmissível
c) Obrigatório	d) Omisso
e) Opcional	f) Não-opcional
g) Deve	h) Deveria
i) Louvável	j) Indiferente / significativo
k) Melhor que / melhor / bom / ruim	l) Reivindicação / liberdade / poder / imunidade
m) Responsabilidade	n) O mínimo que se pode fazer
o) Culpar / elogiar	p) Agência / ação

Fonte: https://plato.stanford.edu/entries/logic-deontic/ (2021).

De um modo geral, a engenharia e a inteligência artificial pesquisam a correspondência de uma estrutura de lógica deôntica com um código moral (premissas ou predições expressas) para agência e automação (Bringsjord, 2018). O foco nas regras lógicas a partir de um código moral aplicado à programação conduz à avaliação de cenários por meio de critérios éticos. Essas regras devem ser suficientemente claras para a agência moral, uma vez que elas são aplicadas na inteligência artificial médica, legal, educacional, automotiva, dentre outras. Além da clareza, as regras devem ter como

características, a confiabilidade, transparência, respeito à autonomia dos sujeitos e responsabilidade. Em relação a essas características, perguntamos se podemos atribuir responsabilidade moral aos agentes inteligentes artificiais?

> O que é necessário e suficiente para que uma entidade artificial seja um agente moral? [...]
> O debate da AMA[6] foi focado principalmente em duas concepções rivais da agência moral humana: a chamada visão padrão e a visão funcionalista, respectivamente. Ambas têm variantes diferentes e, como veremos, pode haver motivos para questionar até que ponto elas realmente conflitam. No entanto, o ponto de partida do debate da AMA é a suposição de que essas duas concepções de agência moral são 1) Incompatíveis e 2) Têm implicações diferentes para a possibilidade de AMA.
> A visão padrão da agência moral humana é que os agentes morais devem atender às condições de racionalidade, livre arbítrio ou autonomia e condições da consciência fenomenal. A visão funcionalista é que a agência requer apenas comportamentos e reações específicas que os defensores da visão padrão visualizassem como meros indicadores das capacidades (Behdadi e Munthe, 2021, pp. 196-197).

Considerando a concepção padrão, perguntamos como é possível atribuir consciência aos agentes inteligentes artificiais? Esses agentes possuem crenças sustentadas por uma consciência fenomenal? O que significa esse tipo de consciência nessa perspectiva? Há uma abrangência de conceitos, problemas, caracterizações, descrições, teorias metafísicas e teorias especificas sobre o que seja a consciência. Para a inteligência artificial, um estado consciente representa um estado fenomenal ligado à "organização espacial, temporal e conceitual da nossa experiência do mundo e de nós mesmos como agentes nele" (Van Gulick, 2014, n.p.). Deixamos claro que esse tipo de concepção é uma variante dentre as diversas tipologias, mas relaciona-se com a visão padrão, uma vez que considera os agentes inteligentes artificiais como uma tecnologia que organiza o espaço, tempo, sintaxe de linguagem e experiência com o mundo. Por sua vez, a atribuição de responsabilidade moral à agência moral artificial vai além da aplicação da deontologia na construção de sistemas tecnológicos inteligentes. Como podemos atribuir responsabilidade moral em agentes artificiais inteligentes se eles não possuem estruturas de sensibilidade, sentimentos e afetos? Nesse aspecto, não compreendemos que essa atribuição seja cabível. A resposta a essa questão envolve considerações metafísicas e ontológicas sobre a correlação entre o real e artificial. Se há uma distinção ontológica entre os artefatos e seres humanos, então apenas atribuímos sentimentos morais aos agentes inteligentes artificiais.

Por sua vez, a visão funcionalista da inteligência artificial considera os agentes como seres sem mente e atuantes por comportamento. A visão funcionalista defende condições como interatividade, independência e adaptabilidade:

[6] AMA: *Artificial Moral Agency* (Tradução: Agência Moral Artificial).

> A visão funcionalista da agência moral foi mais claramente defendida no debate da AMA por Floridi e Sanders, que rejeitam critérios como a consciência, e adotam uma 'moralidade sem mente' (2004 p. 351). Seu ponto de partida é a observação de que as entidades podem ser agentes morais dependendo do nível de abstração escolhido ao inferir critérios gerais de instâncias paradigmáticas da agência moral humana [...]. Floridi e Sanders [...] oferecem o seguinte conjunto de condições para a agência moral:
> 1. Interatividade: E interage com seu ambiente;
> 2. Independência: E tem a capacidade de mudar a si mesmo e suas interações independentemente de influência externa imediata;
> 3. Adaptabilidade: E pode mudar a maneira pela qual 2 é atualizada com base no resultado de 1.
> Embora a condição 1 corresponda aproximadamente à sua contraparte na visão padrão, a condição 2 parte significativamente da visão padrão, não exigindo a presença de estados mentais internos. A condição 3 também é diferente. É mais fraca, pois não exige que as ações de E sejam imediatamente causadas por eventos que se enquadram em 2, mas é mais forte, pois uma condição de capacidade de resposta que vincula 2 e 1 (Behdadi e Munthe, 2021, pp. 198-199).

Adotar uma moralidade "sem mente" implica em um agenciamento caracterizado por comportamento. Considera-se, dentro das dimensões da inteligência artificial, que os agentes inteligentes artificiais simulem o comportamento humano a partir de critérios morais. Nesse aspecto, temos a influência da deontologia, mas sem visar uma racionalidade ideal. Podem existir pontos de contato entre as visões padrão e funcionalista, mas isso corresponde à interatividade com o meio. As condições de independência e adaptabilidade tornam essa agência diferente em termos de mudança de comportamento.

3 ROBÔS INTELIGENTES E ÉTICA DAS MÁQUINAS

Um robô um é artefato físico que se move com certa autonomia. Como todo agente artificial, ele possui atuadores que interagem com o meio, sendo capaz de ter um *feedback* ou não dos dados desse meio. Declaramos que nem toda máquina é um robô e nem todo robô é um agente artificial inteligente. Máquinas podem agir sobre o meio e outros artefatos, mas sua ação não implica em autonomia: um arado puxado por bois atua sobre o meio, mas ele não possui sensores, não é automatizado e nem é programável. De acordo com Muller:

> [...] os robôs são máquinas físicas que se movem. Os robôs estão sujeitos a impactos físicos, normalmente através de "sensores", e exercem força física sobre o mundo,

> normalmente através de "atuadores", como uma pinça ou uma roda giratória[7]. Consequentemente, carros ou aviões autônomos são robôs, e apenas uma minúscula porção de robôs é "humanoide" (em forma humana), como nos filmes. Alguns robôs utilizam IA, outros não: Os robôs industriais típicos seguem cegamente scripts completamente definidos com um mínimo de input sensorial e sem aprendizagem ou raciocínio (cerca de 500.000 novos robôs industriais são instalados todos os anos [...]. É provavelmente justo dizer que, embora os sistemas robóticos causem mais preocupações no público em geral, os sistemas de IA têm maior probabilidade de ter um impacto maior na humanidade. Além disso, os sistemas de IA ou robótica para um conjunto restrito de tarefas têm menos probabilidade de causar novos problemas do que sistemas mais flexíveis e autônomos (Muller, 2020, n.p.).

Há em nossa cultura midiática uma propensão a confundir robôs com androides, humanoides, ciborgues e inteligência artificial. Androides são justamente os robôs humanoides e os ciborgues são a integração corporal e neural de máquinas com os seres humanos (uma perspectiva defende que ciborgues só existirão com a integração do sistema nervoso com artefatos). Não iremos detalhar o que sejam androides e ciborgues, mas sim a relação dos robôs com os sistemas de inteligência artificial e a ética. Há robôs que são criados pela tecnologia da automação para seguir programações simples ou reprogramações mais versáteis. Na indústria automotiva, há robôs colaborativos dotados de sensores especiais que avaliam a qualidade do produto[8] a fim de evitar falhas de pintura, soldagens e componentes mal colocados. Esses robôs são versáteis e reprogramáveis e podem utilizar inteligência artificial para aprender com o meio, processar informações, resolver problemas, tomar decisões e agir a partir de um complexo sistema de *inputs* e *outputs.* Um bom exemplo são os carros autônomos[9] que, através de sistemas de inteligência artificial baseados em redes neurais, tomam decisões complexas no trânsito.

O avanço da tecnologia dos robôs autônomos dotados de inteligência artificial leva a debates éticos ligados ao campo da Ética das Máquinas. Esse campo estuda temas como privacidade e vigilância, interação humano-robô, automação e emprego e sistemas autônomos. Existe uma relação desses temas com a inovação tecnológica, internet das coisas, cidades inteligentes, governança inteligente, robôs de cuidados, robôs sexuais, mudanças da natureza do trabalho e realocação humana para outras atividades, veículos autônomos e armas autônomas (Muller, 2020). Em relação à privacidade e vigilância, podemos considerar o seguinte:

---

[7] O autor refere-se às pinças e rodas automatizadas largamente utilizadas na indústria atual para a fabricação de carros, circuitos eletrônicos etc.

[8] O Snake desenvolvido em parceria pela Empresa Brasileira de Pesquisa e Inovação Industria (EMBRAPII), Instituto Tecnológico da Aeronáutica (ITA) e General Motors (GM) é um exemplo de robô automatizado de forma versátil que avalia a soldagem na indústria veicular. *Cf.*: https://embrapii.org.br/projetos-embrapii/robo-snake/.

[9] Há uma série de projetos de carros autônomos desenvolvidos pela Toyota, Tesla, BMW, Mercedes-Benz etc. O projeto da Waymo já é uma realidade: https://waymo.com/intl/es/waymo-driver/.

> Os dispositivos robóticos ainda não desempenharam um papel importante nesta área, exceto no patrulhamento de segurança, mas isto irá mudar quando forem mais comuns fora dos ambientes industriais. Juntamente com a "internet das coisas", os chamados sistemas "inteligentes" (telefone, TV, forno, lâmpada, assistente virtual, casa...), a "cidade inteligente" [...] e a "governança inteligente", eles estão prestes a fazer parte do mecanismo de coleta de dados que oferece dados mais detalhados, de diversos tipos, em tempo real, com cada vez mais informações (Muller, 2020, n. p.).

Sobre o patrulhamento e vigilância, podemos dizer que encontramos formas "mais gentis" para controlar e manter a ordem social. Michel Foucault (1975) considera que a sociedade moderna criou formas mais suaves de controle: fábricas, hospitais e escolas. Nossas técnicas de manutenção da disciplina têm se aprimorado e não podemos deixar de declarar que a tecnologia tem colaborado bastante com isso. Boa parte do controle do trabalho, disciplina social, ordenamento de hábitos e crescimento das cidades é derivado dos ambientes industriais. Nesse aspecto, podemos citar que determinados modelos de organização, planejamento e administração do trabalho nasceram a partir das indústrias: taylorismo, fordismo e toyotismo. Recentemente, as novas tecnologias da informação e comunicação têm modificado as formas de relacionamento social, familiar e de trabalho. Essas tecnologias popularizaram-se e temos acesso a uma série de dispositivos que colhem informações sobre nossos interesses, gostos e vida pessoal. Utilizamos *smartphones* sincronizados com notebooks, televisões, assistentes de comando por voz, sistemas de navegação espacial, lâmpadas, dentre outros que vão coletando e compartilhando informações: isso é chamado de internet das coisas (IoT).

Ao coletar essas informações, empresas, governos e outras instituições passam a ter conhecimento sobre os indivíduos conectados a esse sistema. Essas conexões permitem, por exemplo, monitorar o trabalho, as condições ambientais de uma fazenda, a entrada e saída de pessoas em espaços públicos e privados, trânsito de veículos através de aplicativos etc. A incorporação dessas tecnologias fornece debates éticos e jurídicos sobre a legalidade em se coletar nossas informações pessoais. Como consequência, as cidades inteligentes têm evoluído a partir do avanço tecnológico da informação e comunicação e do compartilhamento dessas tecnologias com uma parte da população. Uma cidade inteligente pode controlar o fluxo de caminhões de lixo, incluindo rotas mais curtas, desligamento de luzes de postes, monitoramento robótico de câmeras para o trânsito, identificação de pessoas e situações criminosas. Isso permite que os governos tomem decisões baseadas nos dados compartilhados digitalmente a fim de melhorar a administração pública e os serviços aos cidadãos. As críticas aos sistemas inteligentes, cidades inteligentes e governo inteligente voltam-se para as desigualdades de acesso, alfabetização tecnológica e digital: nem todos os indivíduos possuem recursos para ter equipamentos com boa conectividade, qualidade e alcance. A alfabetização tecnológica e digital

consiste não apenas em ter o acesso às tecnologias e a vida digital, mas ter conhecimento básicos de como uma tecnologia funciona, para que serve, suas relações econômicas e políticas.

Um outro ponto importante para o debate da ética das máquinas é a interação humano-robô. Já é um fato nossas interações com a computação afetiva, uma vez que somos atendidos por gravações por assistentes virtuais que simulam sentimentos e emoções através de vozes de boas-vindas, *emojis* e sistemas que parecem diagnosticar o que estamos sentindo. Em termos da robótica, essa interação pode ser observada nos robôs cuidadores:

> A utilização de robôs nos cuidados da saúde humana está atualmente ao nível de estudos de conceito em ambientes reais, mas pode tornar-se uma tecnologia utilizável dentro de alguns anos, e tem levantado uma série de preocupações para um futuro distópico de cuidados desumanizados [...]. Os sistemas atuais incluem robôs que apoiam cuidadores humanos (por exemplo, no levantamento de pacientes ou no transporte de material), robôs que permitem aos pacientes fazerem certas coisas sozinhos (por exemplo, comer com um braço robótico), mas também robôs que são dados aos pacientes como companhia e conforto [...] (Muller, 2020, n.p.).

Os robôs cuidadores podem contribuir para a melhoria do quadro de saúde do paciente e na colaboração com cuidadores humanos. Com o envelhecimento geral da população, os robôs cuidadores serão importantes para o apoio às pessoas com idade mais avançada, mas pode retirar o trabalho de cuidadores humanos. A desumanização, quando adotamos esses robôs, decorre de aspectos como falta de empatia da máquina por mais avançada que seja, atenção, conforto e distanciamento do contato humano. Esse tipo de situação pode provocar mudanças profundas na medicina, enfermagem, psicologia e áreas afins. Em termos éticos, os robôs cuidadores podem gerar uma desvalorização do cuidado e das relações pessoais.

Outro tipo de robôs, os robôs sexuais, podem alterar a forma como nos relacionamos afetivamente e sexualmente. De certa forma, algumas pessoas já utilizam artefatos sexuais para realizarem seus desejos, mas os robôs tendem a criar um conforto:

> Vários otimistas da tecnologia argumentam que os humanos provavelmente estarão interessados em sexo e companhia com robôs e se sentirão confortáveis com a ideia [...]. Dada a variação das preferências sexuais humanas, incluindo brinquedos sexuais e bonecas sexuais, isto parece muito provável: a questão é se tais dispositivos devem ser fabricados e promovidos, e se devem haver limites nesta área delicada. Parece ter entrado na corrente principal da "filosofia do robô" nos últimos tempos [...] (Muller, 2020, n.p.).

Os seres humanos têm o hábito de se relacionar afetivamente com artefatos e objetos. Uma questão ética é se a relação entre robôs e humanos pode ser considerada uma "amizade". Além disso, há críticas quanto à dependência emocional que um ser humano pode criar com um artefato e se os

robôs sexuais venham a ser parte de uma contínua indústria pornográfica e alimentar situações como o trabalho sexual e escravidão (Muller, 2020). Outras críticas realçam a objetificação de nossas relações humanas, degradação moral humana, aumento de problemas psicológicos, irrealismo e ilusão nas relações. Questões de gênero podem criticar que os robôs sexuais tendem a explorar o biótipo feminino em excesso, acentuando desigualdades. Filosoficamente, a interação entre robôs e humanos gera mudanças nas características dos relacionamentos.

Em relação à automação e emprego, ressaltamos que o estudo da inteligência artificial aplicada à robótica é de interesse estratégico para os governos e suas instituições oficiais, é de fundamental importância para manter redes de comunicações, é central nas transações comerciais, econômicas e financeiras entre empresas e serviços, mas, por outro lado, tem sido alvo de debates críticos quanto à substituição ou à realocação de mão de obra humana para outros setores produtivos da economia (Mindell e Reynolds, 2020). O avanço da inteligência artificial é um processo que implica diretamente, na realidade do trabalho. A natureza do trabalho é algo que se altera com as diversas inovações científicas e técnicas que os seres humanos exploram desde antiguidade. De um modo geral e historicamente, as sociedades organizam o trabalho a partir do controle das técnicas agrícolas, de observação do tempo e clima, de conhecimentos escritos, de conhecimentos matemáticos aplicados ao engenho humano, ao domínio da metalurgia, fontes de energia e revolução profunda nos meios de produção. A revolução industrial é o marco de organização do trabalho em torno das inovações científicas e tecnológicas. Há autores que defendem as teses de uma 1ª, 2ª, 3ª e 4ª revoluções industriais. Nesse momento, estaríamos entrando na 4ª revolução industrial devido ao avanço dos sistemas tecnológicos informacionais, comunicacionais, digitais e de inteligência artificial. Em cada uma dessas revoluções, o trabalho mudou em termos de leis, carga horária de trabalho, relações interpessoais, consumo de produtos, meios de transporte, circulação de dinheiro etc. Essas mudanças são tão profundas que impactaram no tempo de trabalho, no local em que se trabalha, na produtividade, na gestão, nas formas de reunião, no despacho de documentos na comunicação administrativa e sua maior digitalização. Como consequência, essas mudanças tão rápidas ainda estão estruturando os modelos educacionais, uma vez que é necessário alfabetizar as pessoas quanto ao uso das tecnologias para a produtividade e geração de riquezas:

> Parece claro que a IA e a robótica levarão a ganhos significativos de produtividade e, portanto, de riqueza geral. A tentativa de aumentar a produtividade tem sido frequentemente uma característica da economia, embora a ênfase no "crescimento" seja um fenômeno moderno [...]. No entanto, os ganhos de produtividade, através da automação, normalmente significam que menos pessoas são necessárias para o mesmo resultado. Contudo, isto não implica necessariamente uma perda global de emprego, porque a riqueza disponível aumenta e isso pode aumentar a procura o suficiente para contrabalançar o ganho de produtividade. A longo prazo, a maior

> produtividade nas sociedades industriais conduziu a mais riqueza em geral. Ocorreram grandes perturbações no mercado de trabalho no passado, por exemplo, a agricultura empregava mais de 60% da força de trabalho na Europa e na América do Norte em 1800, enquanto, em 2010, empregava 5% na UE e ainda menos nos países mais ricos (Comissão Europeia 2013). Nos 20 anos entre 1950 e 1970, o número de trabalhadores agrícolas contratados, no Reino Unido, foi reduzido em 50% [...]. Algumas destas perturbações levam a que indústrias com maior intensidade de mão-de-obra se desloquem para locais com custos de mão-de-obra mais baixos. Este é um processo contínuo (Muller, 2020, n.p.).

Historicamente, o surgimento de novas tecnologias encerrou determinadas profissões, mas, por outro lado, iniciaram outras: o avanço das tecnologias de telecomunicações encerrou a profissão de telefonista, mas, por outro lado, exige uma maior qualificação de técnicos que trabalham com instalação de internet e redes. O avanço da automação, robótica e inteligência artificial parecem substituir o trabalho humano de um modo mais amplo. Embora o trecho acima destaque, historicamente, que uma maior geração de riquezas crie maior procura, isso não é uma declaração que nos dê segurança quanto aos postos de trabalho futuros: seria fundamental apresentar dados econômicos mais precisos sobre os futuros trabalhos e quais as demandas são necessárias para esses trabalhos. No atual momento, o mercado de trabalho relacionado à inteligência artificial e da automação robótica exige a contratação de profissionais altamente qualificados e bem remunerados, enquanto os trabalhos em fábricas e escritórios tendem a diminuir e serem mal remunerados (Muller, 2020).

Por fim, os sistemas autônomos, como os veículos autônomos e armas autônomas, são sistemas mecatrônicos inteligentes que funcionam com bastante independência em relação à intervenção humana. A construção de veículos autônomos visa elaborar sistemas que façam uma condução mais segura. O comportamento humano é imprevisível na direção de veículos e estamos sujeitos a acidentes, uma vez que ocorrem falhas humanas derivadas de imprudência, imperícia e negligência. É comum que o ser humano cometa infrações como dirigir com excesso de velocidade, ultrapassar semáforos e faixas, não dar sinais de mudança de direção etc. A fim de evitar esses problemas, a indústria de veículos autônomos aposta na segurança de seus projetos, porém, caso ocorra um acidente com esta tecnologia, a quem atribuímos a responsabilidade?

Por sua vez, as armas autônomas não são, por exemplo, artefatos simplesmente teleguiados como mísseis e veículos. Eles são independentes e realizam missões complexas (Darpa, 1983). O problema é que a utilização dessas armas pode retirar a responsabilidade humana em guerras e assassinatos, uma vez que é difícil identificar quem ordenou o ataque:

> A assimetria crucial, onde um lado pode matar impunemente e, portanto, tem poucas razões para não o fazer, já existe em guerras convencionais de drones com armas controladas remotamente (por exemplo, os EUA no Paquistão). É fácil imaginar um pequeno drone que procure, identifique e mate um indivíduo humano [...]. Estes são

> os tipos de casos apresentados pela *Campanha para parar os Robôs Assassinos* e outros grupos ativistas. Alguns parecem equivaler a dizer que as armas autônomas são de fato armas..., e as armas matam, mas, ainda assim, as fabricamos em números gigantescos. Em matéria de responsabilização, as armas autônomas podem tornar mais difícil a identificação e a acusação dos agentes responsáveis– mas isto não é claro, dados os registros digitais que se podem manter, pelo menos numa guerra convencional [...] (Muller, 2020, n.p.).

É notável a utilização de drones em guerras entre a Rússia *vs.* Ucrânia, em conflitos entre o Irã *vs.* Israel e em operações antiterroristas dos E.U.A contra o Talibã. O Direito Internacional Humanitário[10] é bem claro quanto ao seu código ético-jurídico em caso de conflitos armados: 1. Princípio da Humanidade: reduzir o sofrimento humano ao decidir pelo ataque menos danoso; 2. Princípio da Necessidade: um ataque militar específico só pode ser realizado após o esgotamento de todas as tentativas possíveis de diplomacia e 3. Princípio da Proporcionalidade: se o alvo de um ataque militar é específico, então não se pode ter uma desproporcionalidade desse ataque aos civis e outras infraestruturas que não são o objetivo da missão. Mediante a esses princípios, notamos que, em combates sem o uso de armas autônomas, há infrações cujo potencial letal é grande. Nesse sentido, questionamos se o uso de armas autônomas pode causar um ataque mais danoso, não reconhecer protocolos diplomáticos e ser desproporcional devido à sofisticação tecnológica. Assim, a questão filosófica que se levanta aqui é em torno da responsabilidade moral dos agentes artificiais inteligentes:

> Neste contexto, responsabilidade implica autonomia, mas não o inverso, pelo que podem existir sistemas que tenham graus de autonomia técnica sem levantar questões de responsabilidade. A noção mais fraca e mais técnica de autonomia na robótica é relativa e gradual: diz-se que um sistema é autônomo em relação ao controle humano até certo ponto. Existe aqui um paralelo com as questões de preconceito e opacidade na IA, uma vez que a autonomia também diz respeito a uma relação de poder: quem está no controle e quem é o responsável? (Muller, 2020, n.p.).

Tendemos a atribuir responsabilidade moral aos robôs na medida que a distância entre nossas características humanas e artificiais diminui, mas isso é um senso comum. Robôs podem ter autonomia, mas essa autonomia é suscetível a uma variação de contextos, graus e controle. Por mais que existam projetos que desenvolvam a aplicação da deontologia em máquinas, a responsabilidade moral é praticamente uma condição ética que depende da consciência:

> Em sua resposta a Wolf, Watson [...] concorda que algumas abordagens à responsabilidade — ou seja, visões de autorrevelação [...] — focam estritamente em se o comportamento é atribuível a um agente. Mas Watson nega que essas atribuições constituam uma forma meramente superficial de avaliação. O comportamento que é

[10] Para maiores detalhes, conferir o site da *Rule of Law in Armed Conflicts da Geneve Academy*: https://www.rulac.org/legal-framework/international-humanitarian-law.

> atribuível a um agente é porque emana de seu sistema de avaliação e, frequentemente, revela algo interpessoal e moralmente significativo sobre a "orientação avaliativa fundamental" do agente [...]. Assim, atribuições de responsabilidade nesse sentido de responsabilidade-como-atribuibilidade são "centrais para a vida ética e avaliação ética" [...] O atribuicionismo assemelha-se às visões de autorrevelação [...] na medida em que ambas focam na maneira como o comportamento de um agente responsável revela características moralmente significativas do *self* do agente (Talbert, 2024, n.p.).

Essas noções de autorrevelação estão fundamentadas na existência de um *self* ou consciência responsiva moralmente significativa. O valor que atribuímos a um comportamento ou código moral não depende exclusivamente de fatores deontológicos, mas há percepções afetivas, emocionais, culturais e estéticas. Entendemos que um robô sofisticado pode tomar decisões morais autonomamente com base em sistemas lógicos, mas isso não fornece decisões significativamente afetivas e estéticas.

CONSIDERAÇÕES FINAIS

Compreendemos que as noções de agência moral entre a inteligência artificial e robôs necessitam serem investigadas filosoficamente. Essa relação nos conduz a entender quais argumentos são utilizados para considerar a inteligência artificial moralmente responsável. Nesse sentido, observamos que a inteligência artificial tende seguir uma racionalidade ideal que visa a construção de sistemas lógicos deônticos ou tende a simular o comportamento humano para a tomada de decisões.

A discussão sobre a atribuição de agência moral à inteligência artificial é importante para os debates filosóficos sobre ética e sua relação com os agentes inteligentes artificiais, algoritmos, aprendizado de máquina e aplicação de lógicas deônticas. Como ponto de partida, consideramos o conceito de Russel e Norvig (2010) sobre agentes inteligentes. É essencial, na história da inteligência artificial, que se entenda a noção de agência, inteligência e como os agentes inteligentes artificiais atuam sobre o meio através de mecanismos de entrada e saída de dados. Russel e Norvig (2010) consideram que os agentes inteligentes artificiais ou seguem uma racionalidade ideal ou seguem um comportamento ideal. Essa racionalidade ideal está voltada para a organização lógica de comandos, fazendo com que os agentes artificiais inteligentes aprendam com o meio e se auto-organizem. Compreender o funcionamento da inteligência artificial é crucial para um senso de que agentes artificiais inteligentes podem tomar decisões morais.

Para Marc Champagne (2021), atribuir responsabilidade moral à inteligência artificial e robôs deve, antes de tudo, responder a um problema metafísico-ontológico: há distinção entre seres humanos e entidades artificiais inteligentes? Apesar da distância entre seres humanos e máquinas estar

diminuindo, não é cabível atribuir um agenciamento moral aos artefatos inteligentes, uma vez que a resposta ao problema metafísico-ontológico não está resolvida.

A relação entre agência moral e a visão padrão e funcionalista nos apresentou a relação da ética com o papel dos algoritmos na tomada de decisões, a aplicação da lógica deôntica à inteligência artificial e as definições do que é uma inteligência artificial voltada para uma racionalidade e outra voltada para o comportamento.

Podemos dizer que há um campo de estudos conhecido como ética dos algoritmos. Nesse campo, estuda-se, por exemplo, o aprendizado de máquina. Angeluci e Rossetti (2021) fornecem argumentos que estabelecem relações entre ética e algoritmos, considerando os padrões estabelecidos nos algoritmos. Segundo Bringsjord (2018), esses padrões devem possuir um código moral que possui uma lógica deôntica: uma lógica voltada para a aplicação de comandos morais. Esse código moral fornece uma estrutura de aplicação para outras como a engenharia e própria inteligência artificial.

No exame da visão padrão, Behdadi e Munthe (2021) argumentam que os agentes inteligentes artificiais devem atender às condições de racionalidade, livre-arbítrio, autonomia e consciência fenomenal. Quanto à noção de racionalidade, ela está de acordo com os pontos de vista de Russel e Norvig (2010), uma vez que é entendido que os agentes inteligentes artificiais estão aprimorando seu processamento de informações e tomada de decisões com base em modelos lógicos. Em relação ao livre-arbítrio e autonomia, devemos perceber que existem graus para a execução de ações, tendo em vista à complexidade de cenários: um veículo autônomo tem um grau de autonomia maior ao lidar com seu próprio sistema, com o usuário e com o meio onde há outros veículos, sinais de trânsito, relevo, navegação, tempo e, principalmente, pessoas – é um sistema complexo. No caso da consciência fenomenal, constatamos que, primeiramente, definir o que é consciência é um profundo problema filosófico, visto que existem uma série de posicionamentos metafísicos, descrições, causas e teorias específicas. Entendemos que a utilização do termo – consciência fenomenal – necessita de maiores detalhes quanto a sua conceituação, estruturação ontológica, aplicação com a inteligência artificial e, principalmente, relação com a uma consciência moralmente responsiva.

O problema de um *self* ou de um eu reflexivo é fundamental para o campo da ética e reponsabilidade moral porque concerne às decisões tomadas pelos seres humanos. Tais decisões partem não só de raciocínios, mas de estruturas afetivas e emocionais ligadas à dinâmica psicológica do indivíduo, mas também ao conjunto de valores significativos compartilhados socialmente e culturalmente. Nesse sentido, precisamos avaliar o quanto uma inteligência artificial será capaz de refletir, avaliar e dar respostas responsáveis quanto à cultura e sociedade.

Nossa análise na seção 3, procurou entender como a ética e a inteligência artificial pode ser aplicada à robótica. Verificamos os argumentos de Muller (2020) e notamos um conjunto grande de

variáveis quando nos referimos aos robôs. Não há uma espécie de robô e nem toda máquina pode ser considerada um robô. Existem robôs automatizados e existem robôs com inteligência artificial e poucos destes robôs são humanoides. Observamos que há uma série de temas éticos quando tratamos de robôs dotados com inteligência artificial: privacidade e vigilância, interação humano-robô, automação e emprego e sistemas autônomos. Esses temas destacam assuntos como controle e ordem social, coleta de dados e informações pessoais, internet das coisas, cidades inteligentes e governança inteligente. Um desses assuntos – interação humano-robô - é bastante polêmico devido aos problemas éticos aplicados à saúde e relacionamentos afetivos. Em relação à automação e emprego, faz-se necessário maiores dados sobre a geração de empregos no futuro, uma vez que Muller (2020) cita a produção de maior riqueza e maior oferta de trabalho: isso não parece ser uma declaração segura. Percebemos também que as revoluções industriais são marcos importantes para a inovação tecnológica e que o atual momento é marcado pelas inovações da inteligência artificial. Por fim, os sistemas autônomos apresentados, demonstram uma grande preocupação, especialmente quando falamos de armas autônomas, tendo em vista seu potencial letal.

Apesar das diferentes conceituações que podemos fazer em relação à responsabilidade moral, agência moral e consciência, este artigo defendeu um posicionamento atribuicionista em relação à responsabilidade moral. Entendemos que o problema metafísico-ontológico entre seres humanos e entidades inteligentes artificiais não está resolvido e que a consciência moral responsiva manifesta uma defesa de um sujeito reflexivo ético. Não sabemos se tal funcionamento ético pode ser aplicado à inteligência artificial e, por isso, apenas atribuímos a ela responsabilidade moral, mas é apenas uma questão de atribuição e não de defesa de que os agentes inteligentes artificiais são moralmente responsáveis.

Artigo recebido em: 16/09/2024
Artigo aceito em: 14/01/2025
Artigo publicado em: 31/03/2025

REFERÊNCIAS

ANGELUCI, Alan; ROSSETTI, Regina. Ética algorítmica: questões e desafios éticos do avanço tecnológico da sociedade da informação. *Galáxia*, São Paulo, n. 46, 2021, p. 1-18. DOI: https://doi.org/10.1590/1982-2553202150301.

MINDELL, David; REYNOLDS, Elisabeth. Inteligência artificial e trabalho: Panorama setorial da internet. [S.l.]: MIT, 2020.

BEHDADI, D.; MUNTHE, C. A normative approach to artificial moral agency. *Minds & Machines*, v. 30, p. 195–218, 2020. DOI: https://doi.org/10.1007/s11023-020-09525-8.

BRINGSJORD, Selmer. Artificial intelligence. In: *Stanford Encyclopedia of Philosophy*. 2018. Disponível em: https://plato.stanford.edu/entries/artificial-intelligence/. Acesso em: 15 abr. 2025.

BRINGSJORD, Selmer. Psychometric artificial intelligence. *Journal of Experimental and Theoretical Artificial Intelligence*, v. 23, n. 3, p. 271–277, 2011. DOI: https://doi.org/10.1080/0952813X.2010.502314.

CHAMPAGNE, Marc. The mandatory ontology of robot responsibility. *Cambridge Quarterly of Healthcare Ethics*, v. 30, n. 3, p. 448–454, 2021. Doi: https://doi.org/10.1017/S0963180120000997.

CUPANI, Alberto. *Filosofia da tecnologia*: um convite. Florianópolis: Editora UFSC, 2016.

DANTAS, Danilo Fraga. Epistemologia Computacional: uma provocação. *Revista Perspectiva Filosófica*, v. 46, n. 02, pp. 189-221, 2019. DOI: https://doi.org/10.51359/2357-9986.2019.248089.

DARPA. *Strategic computing: new-generation computing technology: a strategic plan for its development and application to critical problems in defense*. ADA141982, 28 out. 1983.

DENNET, D. Artificial Intelligence as Philosophy and as Psychology. *In*: RINGLE, Martin. *Philosophical Perspectives in Artificial Intelligence*. New York: Humanities Press and Harvester Press, 1979, pp. 57-80.

EMBRAPII. *Robô Snake promete revolucionar a vistoria de soldas a laser*. Publicado em 27 maio 2019. Disponível em: https://embrapii.org.br/robo-snake-promete-revolucionar-a-vistoria-de-soldas-a-laser/. Acesso em: 23 ago. 2024.

EUROPEAN COMMISSION. How Many People Work in Agriculture in the European Union? An Answer Based on Eurostat Data Sources. *EU Agricultural Economics Briefs*, n. 8, jul. 2013.

FLORIDI, L. *The Philosophy of Information*. Oxford: Oxford Press, 2011.

FOUCAULT, Michel. *Discipline and Punish*. Alan Sheridan (trans.). New York: Pantheon, 1975.

FRANSSEN, Maarten, LOKHORST, Gert-Jan and VAN DE POEL, Ibo. Philosophy of Technology. *In*: ZALTA, Edward N.; NODELMAN, Uri (ed.). *The Stanford Encyclopedia of Philosophy,* Fall 2024. Disponível em: https://plato.stanford.edu/archives/fall2024/entries/technology/. Acesso em: 23 ago. 2024.

KORB, Kevin B. Introduction: machine learning as philosophy of science. *Minds and Machines*, v. 14, p. 433–440, 2004. DOI: https://doi.org/10.1023/B:MIND.0000045986.90956.7f.

MCNAMARA, Paul and VAN DE PUTTE, Frederik. Deontic Logic. *In*: ZALTA, Edward N.; NODELMAN, Uri (ed.) *The Stanford Encyclopedia of Philosophy,* Fall 2022. Disponível em: https://plato.stanford.edu/archives/fall2022/entries/logic-deontic/. Acesso em: 23 ago. 2024.

MITTELSTADT, B. D. *et al.* The ethics of algorithms: mapping the debate. *Big Data & Society*, v. 3, n. 2, p. 1–21, jul./dez. 2016. DOI: https://doi.org/10.1177/2053951716679679.

MÜLLER, Vincent C., Ethics of Artificial Intelligence and Robotics. *In*: ZALTA, Edward N.; NODELMAN, Uri (ed.). *The Stanford Encyclopedia of Philosophy*. Fall 2023. Disponível em: https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/. Acesso em: 23 ago. 2024.

NEWELL, N. You can't play 20 questions with nature and win: projective comments on the papers in this symposium. *In*: CHASE, W. (ed.). *Visual information processing*. New York, NY: Academic Press, 1973, pp. 1-26.
O' NEIL, Cathy. *Algoritmos de Destruição em Massa*. [s/l]: Editora Rua do Sabão, 2020.
O' NEIL, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing Group, 2017.
RULAC. Rule of Law in Armed Conflicts. *Legal framework: International humanitarian law*. 2017. Disponível: https://www.rulac.org/legal-framework/international-humanitarian-law. Acesso em: 23 ago. 2024.
RUSSELL, S., & NORVIG, P. *Artificial Intelligence*: a modern approach. 3rd edition**,** Saddle River, NJ: Prentice Hall, 2010.
SEARLE, J., *The Mystery of Consciousness.* London: Granta Books 1997.
TALBERT, Matthew. Moral Responsibility. *In*: ZALTA, Edward N.; NODELMAN, Uri (ed.). *The Stanford Encyclopedia of Philosophy*, Fall 2024. Disponível em: https://plato.stanford.edu/archives/fall2024/entries/moral-responsibility/. Acesso em: 23 ago. 2024.
TEIXEIRA, João Fernandes. *Inteligência Artificial*: uma odisseia da mente. São Paulo: Paulus, 2014.
TRIPICCHIO, Adalberto e TRIPICCHIO, Anna Cecília. *Teorias da Mente*. São Paulo: Teccmed Editora, 2003.
TURING, Alan M. Computing machinery and intelligence. *Mind*, v. 59, n. 236, p. 433–460, out. 1950. DOI: https://doi.org/10.1093/mind/LIX.236.433.
VAN GULICK, Robert. Consciousness. *In*: ZALTA, Edward N.; NODELMAN, Uri (ed.). *The Stanford Encyclopedia of Philosophy*, Winter 2022. Disponível em: https://plato.stanford.edu/archives/win2022/entries/consciousness/. Acesso em: 23 ago. 2024.
WAYMO. *Waymo Driver*. Disponível em: https://waymo.com/intl/es/waymo-driver/. Acesso em: 23 ago. 2024.