UNIVERSIDADE DE CAXIAS DO SUL ÁREA DE HUMANIDADES PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA

ANDERSON DE CAMPOS TRIACCA

HEIDEGGER E OS LIMITES DAS A.I.:
ESTUDANDO A POSSÍVEL MORALIDADE DAS MÁQUINAS

CAXIAS DO SUL

2020

ANDERSON DE CAMPOS TRIACCA

HEIDEGGER E OS LIMITES DAS A.I.:
ESTUDANDO A POSSÍVEL MORALIDADE DAS MÁQUINAS

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre em Filosofia, da Universidade de Caxias do Sul, pelo Programa de Pós-Graduação em Filosofia

Orientador: Prof. Dr. Itamar Soares Veiga

CAXIAS DO SUL

2020

Dados Internacionais de Catalogação na Publicação (CIP)
Universidade de Caxias do Sul
Sistema de Bibliotecas UCS - Processamento Técnico

T819h Triacca, Anderson de Campos
Heidegger e os limites das A.I. [recurso eletrônico] : estudando a possível moralidade das máquinas / Anderson de Campos Triacca. – 2020.
Dados eletrônicos.

Dissertação (Mestrado) - Universidade de Caxias do Sul, Programa de Pós-Graduação em Filosofia, 2020.
Orientação: Itamar Soares Veiga.
Modo de acesso: World Wide Web
Disponível em: https://repositorio.ucs.br

1. Inteligência artificial. 2. Filosofia. 3. Ética. 4. Heidegger, Martin, 1889-1976. 5. Searle, John R., 1932 -. I. Veiga, Itamar Soares, orient. II. Título.

CDU 2. ed.: 004.8

Catalogação na fonte elaborada pela(o) bibliotecária(o)
Carolina Machado Quadros - CRB 10/2236

"Heidegger e os limites das A.I.: estudando a possível moralidade das máquinas"

Anderson de Campos Triacca

Dissertação de Mestrado submetida à Banca Examinadora designada pela Coordenação do Programa de Pós-Graduação em Filosofia da Universidade de Caxias do Sul, como parte dos requisitos necessários para a obtenção do título de Mestre em Filosofia. Linha de Pesquisa: Problemas Interdisciplinares de Ética.

Caxias do Sul, 16 de novembro de 2020.

Banca Examinadora:

Participação por videoconferência
Prof. Dr. Itamar Soares Veiga (orientador)
Universidade de Caxias do Sul

Participação por videoconferência
Prof. Dr. Matheus de Mesquita Silveira
Universidade de Caxias do Sul

Participação por videoconferência
Prof. Dr. Róbson Ramos dos Reis
Universidade Federal de Santa Maria

“O exato das ciências exatas não pode ser exatamente determinado, ou seja, por vias de cálculo, mas apenas ontologicamente, e assim sendo o tipo de verdade que se aplica à ciência é no sentido da ciência natural exata. Sua verdade é verificada pela eficiência de seus resultados. Quando este modo de pensar determina o conceito científico de ser humano e o investiga a partir do modelo de conjunto de leis, como agora acontece na cibernética, em pouco tempo a destruição do ser humano será perfeita.”.

Martin Heidegger

RESUMO

A presente dissertação estuda a possibilidade de agentes artificiais ter acesso e fazer uso de algum tipo de moralidade. Apoiando-se na filosofia de Heidegger e Searle, buscamos os caminhos que levaram a ciência até a concepção de agentes artificiais, e veremos como tais desenvolvimentos encontram-se relacionados de maneira muito estreita ao problema do *esquecimento do ser*. A partir do estudo dos paradigmas hegemônicos da Inteligência Artificial tentaremos trazer ao plano do evidente seus pressupostos filosófico-ontológicos, de modo que se torne possível delimitar, ao menos de forma aproximada, até onde sua inteligência é capaz de chegar. Ao final veremos como as máquinas desenvolvidas até o momento não foram ainda capazes de vencer as críticas impostas por Searle e Heidegger, mas que isso não impediu a concepção, mesmo que teórica, de Agentes Morais Artificiais.

Palavras-chave: Agentes Morais Artificiais, AMA, Inteligências de Máquina, Heidegger, Searle.

ABSTRACT

This dissertation studies the possibility of artificial agents to access and make use of some type of morality. Based on the philosophy of Heidegger and Searle, we look for the paths that led science to the conception of artificial agents, and we will see how such developments are very closely related to the problems treated on Being and Time. From the study of the hegemonic paradigms of Artificial Intelligence, we will try to find its philosophical-ontological assumptions, so that it becomes possible to delimit, at least in an approximate way, how can machines access the morality. In the end, we will see how the machines developed so far have not yet been able to overcome the criticisms imposed by Searle and Heidegger, but this did not prevent the conception, even if theoretical, of Artificial Moral Agents (AMAs).

Keywords: Artificial Moral Agents, AMAs, Artificial Inteligence, Heidegger, Searle.

.

SUMÁRIO

INTRODUÇÃO

A evolução das tecnologias nos condicionou a uma realidade onde cada vez mais as *inteligências de máquina* estão tomando o lugar dos agentes humanos em tarefas automatizáveis. A gramática deste texto por exemplo é corrigida por um algoritmo, desenvolvido no Brasil em 1993[1] e comprado pela Microsoft para ser incorporado ao pacote Office. Historicamente falando, isso representa uma das primeiras interações de humanos com inteligências de máquinas, ou melhor dizendo, de inteligências de máquina projetadas para auxiliar tarefas executadas por humanos num âmbito fora de centros científicos. Essas interações evoluíram bastante desde 1993 e se tornaram comuns e rotineiras. Tão comuns ao ponto de nem nos darmos conta de que escondidas nesta aparente banalidade utilitária encontram-se questões filosóficas profundas.

Desde os primórdios da mecânica, as máquinas foram pensadas como auxiliares e extensores do trabalho humano físico, permitindo ao homem executar tarefas que antes eram impossibilitadas por nossas limitações biológicas. A emergência das tecnologias da informação reproduziram o mesmo fenômeno no séc. XX, porém buscando automatizar os trabalhos intelectuais, permitindo executar tarefas que as limitações da nossa mente antes não permitiam (pelo menos não com a mesma eficiência). As máquinas são desde sempre projetadas para interagir com os humanos, e tais interações estão nos obrigando a rever nossos conceitos de *agente*, *inteligência* e *responsabilidade moral*, pois as antigas categorias não acomodam mais um mundo de agentes artificiais em constante autotransformação e aprendizado. Tais agentes serão analisados a partir das obras de alguns teóricos da computação do séc. XX[2].

---

[1]Criado no ICMC-USP. Wakka, Wagner. Canaltech: Corretor ortográfico do Word foi inventado por brasileiros, c2018. Página inicial. Disponível em: <https://canaltech.com.br/curiosidades/voce-sabia-corretor-ortografico-do-word-foi-inventado-por-brasileiros-115116 >. Acesso em: 12 de jul. de 2020.

[2] Alan Turin, John Searle, Allen Newell, John McCarthy e Herbert Simon.

Para Heidegger, nosso modo de acesso ao mundo dá-se pela manualidade (*Zuhandenheit*), ou seja, o acesso aos entes que se encontram a nosso alcance. Através de nossa relação utilitária com eles é que nosso mundo, compreendido como horizonte de sentido, se forma, de modo que tal relação é de caráter constituinte de nossos modos de ser, e é através desta mesma manualidade que o Dasein experiencia originalmente o conhecimento do mundo.

No capítulo I trataremos da visão de Heidegger sobre a técnica, eixo de temático de sua obra a partir do período conhecido como *vira-volta* (*die Kehre*). onde explora a mudança da essência da verdade efetivada pela tradição filosófica, que modificou o conceito hegemônico dos gregos antigos, que era o de *verdade como ἀλήθεια* para o que Heidegger chamou de *verdade como adequação*. Faremos o caminho do *esquecimento do Ser* até a *questão da técnica*, para fundamentar filosoficamente toda base conceitual necessária para compreendermos as atuais capacidades e limitações dos tipos de inteligências de máquina disponíveis. Veremos que os critérios utilizados para imputação de responsabilidade moral nos seres humanos (intencionalidade, motivação, racionalidade) não se aplicam às ações das máquinas, pois suas inteligências são bastante limitadas e específicas, ao contrário da inteligência humana que é capaz de aprendizados abstratos complexos.

No capítulo II consultaremos a história e a literatura técnica da computação, utilizando os *paradigmas simbolista* e *conexionista* como ponto de contato com as teorias filosóficas. Tentaremos trazer ao plano do evidente os pressupostos ontológicos que permitiram o desenvolvimento das inteligências de máquinas com as quais convivemos hoje em dia. Somente compreendendo esses pressupostos é possível apontar com uma certa margem de segurança as limitações de tais inteligências. Ao final do capítulo falaremos dos períodos conhecidos como *AI Winter*, concluindo que os mesmos possuem uma base que pode ser parcialmente compreendida a partir dos conceitos Heideggerianos de *hegemonia científica da técnica*, *esquecimento do ser* e *verdade como adequação*. Tais conceitos nos ajudam a fundamentar a natureza das ações dos agentes artificiais, e a partir deles é possível estabelecer critérios sobre quais deles podem, em alguma hipótese, vir a ser considerados *agentes morais*.

No capítulo III a pesquisa fica centrada na demarcação e identificação da possibilidade de enquadramento dos agentes artificiais dentro do grupo de *agentes morais*. A partir das críticas de Searle e Heidegger, apontamos a impossibilidade das máquinas ter uma moralidade equivalente à humana, principalmente devido ao fato das *inteligências de máquina* não possuirem intencionalidade e nem compreensão semântica das sentenças linguísticas, mesmo que se tornem capazes de imitar com perfeição nossas faculdades comunicativas. Independentemente de nossa compreensão teórica sobre agentes artificiais, a nossa realidade já nos coloca em convivência com *inteligências de máquina* que invariavelmente precisam tomar decisões morais no seu contato com os outros entes do mundo. Agentes que são colocados perante tais situações são tecnicamente chamados de *agentes morais artificiais* (AMAs), e possuem critérios de demarcação que os categorizam de tal forma.

Nem os mais avançados entes artificiais até o momento foram capazes de agência ao ponto de ser passíveis de imputação de responsabilidade moral, porém isso não significa que os mesmos não precisem ter acesso a algum tipo de moralidade que modere seu comportamento para com os demais entes do mundo. Diante do contexto desses capítulos buscaremos resolver o seguinte problema: até que ponto os agentes artificiais podem acessar e fazer uso da moralidade?

1 - A τέχνη MODERNA COMO NOVA ESSÊNCIA DA VERDADE

As primeiras compreensões de homem nos transportam até a antiguidade clássica, e apontam justamente para sua racionalidade como diferença específica característica da categoria *humano*. Desde então a tradição filosófica o interpretou como superior no que diz respeito a seu λόγος discursivo[3], e humanidade acabou se tornando sinônimo de racionalidade, ao mesmo passo que racionalidade tornou-se sinônimo de inteligência. Tal confusão terminológica não é casual, mas reflexo do direcionamento que a civilização ocidental tomou a partir da Grécia antiga. A tecnologia tornou a compreensão do conceito de *inteligência* problemática: do ponto de vista da epistemologia, pela primeira vez temos que lidar com modelos computacionais (*algoritmos*) que, dentro de sua realidade delimitada, são também universais e necessários, indicando que o que compreendemos por inteligência não é algo tão exclusivo do homem, nem sequer dos entes biológicos.

Tais modelos computacionais representam o ponto mais alto das tecnologias desenvolvidas a partir do florescimento científico do séc. XVIII, que foram evoluindo até a criação da informática e tecnologias da informação no séc. XX. Cientificamente falando, este florescimento é resultado de um projeto de dominação da natureza inaugurado por Galileu e Newton, e teve sua fundamentação epistemológica com Kant, que separou *ciência* e *metafísica*, abrindo os caminhos teóricos para o desenvolvimento da *ciência* como *técnica*. As revoluções industriais levaram adiante tal projeto a um nível em que a objetificação da natureza se tornou paradigma hegemônico das ciências em geral, e por consequência, orientadora da visão de mundo da sociedade globalizada, que passou a compreendê-la como indicadora de progresso humano.

> A objetivação da natureza assim determinada é por ela mesma o projeto de natureza enquanto vista de um âmbito objetual que pode ser dominado. Os passos decisivos no desdobramento deste projeto de uma natureza dominável foram levados a cabo por Galileu e Newton. O que lhes importa é como a natureza é representada, e não o que ela é. (HEIDEGGER, 2013, p. 213).

[3] "A história da lógica no ocidente, assim como, a partir daí, as ciências das línguas em geral foram determinadas pela teoria grega do logos no sentido da proposição enunciativa." (HEIDEGGER, 2003a, p. 347).

Característica principal de tal projeto é o entendimento da natureza a partir de suas representações, ao invés de uma visão fenomenológica que busca desvelar o que se esconde por detrás dos fenômenos. Pela forte dependência de tais representações, tal paradigma acaba sendo chamado de Paradigma Representacional.

A hegemonia de tal modelo cientifizador resultou numa realidade técnica onde o convívio com *inteligências de máquina* se tornou comum e rotineiro, ao e mesmo passo que nossas inteligências artificiais não conseguem avançar para muito além do âmbito da mera representação. Nossos smartphones nos indicam amizades (e nos enchem de propaganda) através de algoritmos que ficam minerando informações sobre nossos interesses de forma automatizada. Em alguns países já é realidade o uso de carros sem motorista, que interpretam as imagens do mundo ao seu redor e tomam decisões baseadas nelas. Já existem estudos especulativos que tratam da possiblidade das desigualdades sociais estarem sendo acentuadas pela atividade de tais algoritmos[4], dentro outros muitos exemplos que a tecnologia atual nos mostra a cada dia.

A emergência de tais entes artificiais, que de *entes meramente dados* passam a *agentes* (pois de alguma forma tomam algum tipo de decisão baseada em possíveis razões e que resultam em ações no mundo) abre espaço para a possibilidade de ocorrência de incidentes que podem ter implicações morais e legais (acidentes com carros automáticos, mortes em decorrência de cirurgia robotizada, etc.). No caso de ocorrências desta natureza, entra em jogo a discussão sobre a *responsabilidade moral* de seus agentes para que se possa inferir suas implicações morais e jurídicas. Pela primeira vez na história seus agentes são artificiais, resultando em uma brecha conceitual em nossos sistemas, que possui sua raiz no problema ontológico do *esquecimento do ser*.

Entes artificiais baseados em *inteligências de máquina* também possuem, até certo ponto, capacidade de ação, que por sua vez independe de consciência. Para direcionar o texto até tal argumento se fazem necessários dois passos principais: distinguir inteligências humanas de inteligências de

[4] O'NEIL, Cathy. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown, 2016. ISBN 9780553418811.

máquina (que será feito através da discussão entre *Conexionismo* e *Simbolismo*), e distinguir consciência de inteligência, a partir da posição de Searle a respeito das propriedades Sintáticas e Semânticas da mente.

Para realizar o primeiro passo com sucesso, será necessário compreender a forma como a τέχνη vem, ao longo de sua evolução, modificando o status ontológico de certos entes do mundo percebido. Para isso consultaremos as obras de Heidegger em sua segunda fase, onde o conceito de verdade como presença é acusado de ter cometido uma mudança da essência da verdade *aléthica* para uma verdade como adequação. Desta forma, o humanismo teria levado a *Metafísica* ocidental a sua plenitude nos configurando em civilização técnica.

> O desvelamento deve ser arrancado ao velamento, de certo modo deve ser roubado. E, visto que para os gregos, inicialmente, o velamento perpassa e domina a essência do ser como um velar-se, determinando, assim, também o ente em sua presença e acessibilidade("verdade"), por isto a palavra que os gregos usam para aquilo que os romanos chamam de "veritas" e nós chamamos de "verdade", vem caracterizada pelo α-privativo (α-λήθεια).Verdade significa, de início, aquilo que foi arrancado ao velamento. Verdade é portanto, esta conquista pela luta, a cada vez sob a forma do desencobrimento." (HEIDEGGER, 2008, p. 235).

Este estudo será feito na primeira parte deste capítulo, e será precedido de um estudo da discussão *conexionismo* x *simbolismo*, onde os desvelamentos técnicos previstos por Heidegger começaram a se materializar,e que culminará na criação dos cérebros eletrônicos (computadores – automatizadores de trabalho intelectual). Neste ponto faremos uma convergência das críticas dos períodos chamados de*AI Winter*[5] (que são em última análise críticas ao paradigma simbolista esgotado em sua representacionalidade), com as críticas de Heidegger a hegemonia científica dos modelos representacionais, que seriam mero reflexo da mudança na *essência da verdade* e do *esquecimento do ser* a partir da configuração da civilização ocidental como civilização técnica.

[5] AI Winter é o termo utilizado para se referir a períodos onde os investimentos públicos e privados nas pesquisas em AI foram reduzidos, fazendo com que a evolução tecnológica ficasse estagnada. Tal estagnação leva a uma descrença generalizada na tecnologia, que geralmente acaba sendo embasada pela mídia.

O segundo passo será tema do próximo capítulo, onde analisaremos a visão Fenomenológico-Hermenêutica de Heidegger sobre a consciência, compreendendo a mesma como uma propriedade relacional e intencional dos entes, e que se torna problemática quando pensada no contexto do *machine learning*. Finalizado este percurso teremos deixado pronto o campo conceitual necessário para darmos prosseguimento ao estudo da possibilidade da responsabilidade moral a partir dos contextos dos agentes artificiais aqui delimitados.

Historicamente falando, Heidegger viveu as duas primeiras revoluções industriais e viu o mundo agrário tornar-se mecanizado, apontando como característica fundante de nossa época a crescente automatização dos processos *poiéticos*. Os processos produtivos migraram pro contexto da grande fábrica, e tal posição histórica talvez nos ajude a compreender sua postura anti-modernista: os comportamentos humanos são fundados na compreensão utilitária dos entes, e acabam cada vez mais influenciados pelas nossas relações de manualidade, de modo que a compreensão de mundo do Dasein acaba diretamente modificada na medida em que ele se coloca em relação com a τέχνη, através da manipulação de suas coisificações.

> O ente sustentado na posição prévia, por exemplo, o martelo, de início, está à mão como instrumento. Se ele se torna “objeto” de uma proposição ,já se realiza previamente com a sentença proposicional uma mudança na posição prévia. Aquilo com que lidava manualmente o fazer, isto é,a execução, torna-se aquilo “sobre” o que a proposição demonstra. (HEIDEGGER, 1989, p. 215).

Tal momento histórico representaria o ponto mais alto de uma metafísica produtivista inaugurada por Platão e pela filosofia grega, que teria ao longo do tempo culminado no que chamou de *esquecimento do ser*, e que apresenta seu acabamento final na técnica moderna. Sua argumentação se apresenta a partir de *Ser e Tempo*, onde através de uma análise filológica, volta a raiz grega da palavra *fenômeno*, que deriva do verbo φαινεσθαι: mostrar, surgir. φαινομενον é seu modo nominativo, que significa “aquele que se mostra”, ou o “aquilo que φαινω”: auto-mostrante. Logo, o sentido que é resgatado significa “o que se mostra em si mesmo”. Para os gregos, φαινομενον designa simplesmente a totalidade daquilo que se mostra, que também já foi identificado como τά οντά

(aquele que ov, o *ente*). Porém, dentro da analítica existencial de Heidegger, o *ente* pode se mostrar a partir de si mesmo de diversos modos, sendo que inclusive o *ente* pode se mostrar como algo que ele não é em si mesmo. Assim o *ente* não mostra o que ele é, mas apenas o que ele aparenta. Deste modo, *fenômeno* possui dois sentidos: “o que se mostra” e “aparência”.

> A expressão grega *φαινομενον*, à qual remonta o termo “fenômeno”, deriva do verbo *φαινεσθαι*, que significa mostrar-se; *φαινεσθαι* significa, portanto, o que se mostra, o se-mostrante, o manifesto [was sich zeigt, das Sichzeigende, das fenbare].*φαινομενονé*, ele mesmo, uma formação média de *φαινω* - trazer à luz do dia, pôr em claro; *φαινω* pertence à raiz *φα* - como *φω*, a luz, o claro, isto é, aquilo em que algo pode se tornar manifesto, pode ficar visível em si mesmo. Como significação da expressão “fenômeno” deve-se portanto reter firmemente: o-que-se-mostra-em-si-mesmo, o manifesto. Os “fenômenos” são então o conjunto do que está à luz do dia ou que pode ser posto em claro, aquilo que os gregos às vezes identificaram simplesmente com o ente. Ora, o ente pode se mostrar, a partir de si mesmo, de diversos modos, cada vez segundo o modo-de-acesso a ele. Há mesmo a possibilidade de que o ente se mostre como o que ele não é em si mesmo. Nesse mostrar-se, o ente aparenta, ele “é como se...” (HEIDEGGER, 2005, p. 58).

Na terminologia de Heidegger, ficamos apenas com um dos sentidos, o sentido de “aquele que se mostra”, legando para aparência outro sentido e utilização. Assim sendo, acaba por diferenciar o conceito f*enomenológico de fenômeno* do conceito “vulgar” de fenômeno. O conceito “vulgar” corresponderia ao conceito kantiano, onde os fenômenos são a totalidade daquilo que afeta nossa intuição empírica. Para Kant, não temos acesso às *coisas em si* (*nôumenos*), mas somente as *representações* delas servidas por nossa sensibilidade: a estas representações Kant chama *fenômeno*.

Esta nova forma de ver os fenômenos, demonstrada em Ser e Tempo, apontava para uma nova compreensão do conceito de *verdade,* que seria devidamente trabalhada em breve. Tal estudo inaugura a segunda fase de seu pensamento, na preleção *A doutrina platônica da verdade* [Platons Lehre von der Wahrheit], onde afirma que a tomada da metafísica como fundamento do mundo por parte da tradição grega culminou em uma nova compreensão do conceito de *verdade*.

O fenômeno, compreendido fenomenologicamente como *aquilo que se mostra em si mesmo*, implica que a *verdade* não pode mais ser condicionada

as aparências, mas precisa estar de acordo com os fenômenos tal qual se mostram na abertura do Dasein e ao seu *modo-de-acesso*, obedecendo ao método fenomenológico para evitar cair nos vícios da tradição. Ao traduzir diretamente do grego o termo *ἀλήθεια*, aponta o mesmo como origem da essência da verdade como encobrimento/desencobrimento, alinhada com o conceito *fenomenológico de fenômeno*. Porém a tradição metafísica, ao *entificar* o ser, teria substituído o conceito de fenômeno pela mera aparência, e por consequência o conceito da verdade como desvelamento pelo de verdade como adequação e conformidade. Esta mudança de paradigma no que determina a essência da verdade é o que Heidegger chamou de o "que permanece não formulado" na doutrina platônica.

> A doutrina de um pensador é o que permaneceunão formulado nas suas palavras, mas ao qual o homem está aberto, "exposto" [...]. Se quisermos extrair e também conhecer o que um pensador não disse, qualquer que seja a natureza, é necessário considerar o que ele disse [...]. O que permaneceu nãoformulado é um movimento concernente àdeterminação daessência da verdade." (HEIDEGGER, 2006, p. 427).

Tal mudança inaugura uma nova forma de compreender a Φύσις, que seria contrária a visão dos pré-socráticos (Heráclito e Parmênides) de uma presença que surge no desvelar dos fenômenos (diretamente ligada a ideia de *devir*). Esta nova visão começa a questionar pelo ser que se aparenta, e não mais pelo que se desvela, colocando a substância (agora compreendida como οὐσία) no fundamento geral do ser dos entes, e tendo como implicação direta a cristalização do sentido de verdade como adequação entre o percebido e o perceber. A verdade se torna então adequação a um modelo ideal, ao mesmo tempo humano e metafísico, que se desenvolveu até chegarmos na modernidade, onde a essência da verdade se encontra submetida a instrumentalização dos entes em seu contexto utilitário.

Assim sendo, a verdade do modo de desvelamento da técnica acaba por comandar a essência do real na modernidade, culminando no ponto máximo da metafísica humanista do esquecimento do ser. Sendo a técnica uma ἀλήθεια que se dá na ποίησις, e sendo essa ποίησις baseada em finalidade funcional.

Tal cenário torna real o teste de Turing, onde nós conversamos virtualmente com *bots* de atendimento de empresas, cuja finalidade funcional é tentar nos enganar se passando por um atendente humano. Atualmente a

tecnologia acessível aos empresários ainda não chega nem perto de enganar uma inteligência humana, mas será que logo não chegará? E se chegar, teremos elevado o status ontológico do software ao da consciência? Ou apenas teremos reduzido o conceito de humano para “ente que passou no teste de Turing”?

Tais perguntas partem de um pressuposto behaviorista: como fica nossa compreensão de inteligência se um dia uma máquina conseguir imitar o comportamento consciente de um humano? A seguir estudaremos de forma mais aprofundada tais concepções de entes artificiais, onde uniremos *conexionismo* e *simbolismo* em busca da melhor compreensão possível sobre o funcionamento das diversas inteligências de máquina. A compreensão das matrizes epistemológicas de cada uma delas a partir de seus fundamentos teóricos nos dará uma visão abrangente, que será necessária para o estudo que pretendo empreender no capítulo seguinte.

2 - INTELIGÊNCIAS DE MÁQUINA E A TRADIÇÃO FILOSÓFICA

Os teóricos das inteligências de máquina talvez tenham sido os primeiros acadêmicos a tentar unir os conceitos das ciências humanas com as possibilidades de testes empíricos inauguradas pelos avanços tecnológicos das revoluções industriais. Uma parte deles buscará apoio teórico na Psicologia, e veremos que diferentes linhas de pensamento psicológico derivam diferentes formas de compreender as inteligências de máquina. Outra parte buscará apoio mais direto na fisiologia, e todas elas ao se deparar com conceitos tais como *inteligência*, *consciência* e *ação* se viram obrigadas a beber dos autores da filosofia, que desde a antiguidade já se debruçavam sobre os mesmos temas.

Faz sentido se perguntar pelo significado do que entendemos por inteligência? Turing defende que a busca de uma resposta nos obriga a modificar a pergunta. Seu posicionamento o conduza uma alternativa behaviorista: se algo se comportar como uma inteligência, então é inteligente. Porém esta não é a única e nem hegemônica forma de compreender a inteligência dentro da história da computação e das ciências no geral.

> “Given the table corresponding to a discrete-state machine it is possible to predictwhat it will do. There is no reason why this calculation should not be carried out bymeans of a digital computer. Provided it could be carried out sufficiently quickly thedigital computer could mimic the behavior of any discrete-state machine.” (TURING, 1950, p. 440).

Definições estritas de *inteligência* envolvem conceitos como *razão e pensamento*, porém quando lidamos com máquinas essas propriedades se tornam confusas e embaralham nossas categorias humanas. Podemos afirmar que quando um algoritmo de *machine learning* classifica dados está de fato pensando? Tentando responder tal pergunta, Turing afirma que precisamos recorrer ao conceito de *inteligência*, que dentro de sua visão behaviorista seria compreendida como manipulação de símbolos, fazendo da possibilidade de criação de uma máquina capaz de pensar conceitualmente válida. Iremos estudar a posição simbolista original de Turing, a réplica de Searle (a partir do *naturalismo biológico*), e as teorias conexionistas das redes neurais. Cada uma destas linhas de pensamento vai dar origem a diferentes algoritmos com diferentes níveis de autonomia, que posteriormente serão analisados a partir

da base conceitual desenvolvida.

2.1 – A posição simbolista original: Alan Turing

Simbolismo é o nome dado ao primeiro paradigma científico da história da AI. Seu nome diz respeito a teoria de que inteligência é nada mais que manipulação de símbolos, e que máquinas capazes de manipular símbolos são máquinas inteligentes. O grande desafio seria apenas construir máquinas capazes de reproduzir os estados mentais humanos no que diz respeito a sua capacidade de manipulação simbólica.

Tal posição nasce junto com a história da computação através da obra do britânico Alan Turing, que idealizou uma máquina abstrata capaz de computação de dados através da manipulação de símbolos[6]. O projeto acabou se tornando a base de toda computação, e a partir da evolução tecnológica da eletrônica digital, impulsionada pela criação dos transistores e manipulação do silício, novos paradigmas se desenvolveram e hardwares mais poderosos possibilitaram a criação de novas formas de processar e compreender dados computacionais.

Em seu artigo seminal “Computing Machinery and Intelligence” (TURING, 1950), propõe que a pergunta sobre a possibilidade das máquinas pensar seja substituída por um jogo, chamado por ele de *jogo da imitação*. O objetivo é fazer com que uma máquina engane um ser humano num jogo de perguntas e respostas, tentando se passar por humana ao responder perguntas do interlocutor. Se a máquina se comportar (imitar corretamente) um humano, Turing afirma que então podemos considerar a máquina inteligente. Tal posição foi criticada por alguns e defendida por outros, e de certa forma a história da inteligência artificial é uma espécie de movimento em torno de tentar validar ou desvalidar tal teoria fundamental a respeito da inteligência como manipulação de símbolos.

---

[6]“Turing machines, first described by Alan Turing in Turing 1936–7, are simple abstract computational devices intended to help investigate the extent and limitations of what can be computed”. Stanford Encyclopedia of Philosophy: Turing Machines, c2018. Página inicial. Disponível em: <https://plato.stanford.edu/entries/turing-machine>. Acesso em: 10 de ago. de 2020.

> (...)it is difficult to escape theconclusion that the meaning and the answer to the question, "Can machines think?"is to be sought in a statistical survey such as a Gallup poll. But this is absurd.Instead of attempting such a definition I shall replace the question by another, whichis closely related to it and is expressed in relatively unambiguous words. (...)We now ask the question, "What will happen when a machine takes the part of A inthis game?" Will the interrogator decide wrongly as often when the game is playedlike this as he does when the game is played between a man and a woman? Thesequestions replace our original, ‘Can machines think?”. (TURING, 1950, p. 433).

O behaviorismo foi uma corrente psicológica bastante popular no séc. XX, que correspondia a uma tentativa de compreensão da mente através do estudo do comportamento, pois ele era o único aspecto empiricamente observável da mente (dada a tecnologia disponível na época). Para que a Psicologia pudesse ser reconhecida como ciência, ela precisou responder ao que os teóricos chamaram de *vetos kantianos*. Um dos famosos vetos dizia respeito ao *objeto de estudo* de uma Psicologia que se pretendia científica, pois o fenômeno psíquico não era fisicamente observável. Partindo dessa premissa, os pioneiros do *behaviorismo* resolveram estudar o pouco fisicamente observável da psique: o comportamento. Através de seu estudo conseguiram avanços significativos nas teorias dacognição, de modo que agora a ciência tinha certeza que pelo menos uma parte do nosso aprendizado ocorria através de imitação e condicionamento comportamentais, além do desenvolvimento de algumas partes da nossa estrutura mental depender diretamente dos mesmos fatores. Filosoficamente falando, o behaviorismo se alinha com as teorias materialistas, e o talvez o conceito unificador da Psicologia e da Filosofia no que diz respeito a tal posição seja o de *tábula rasa*.

Os psiquiatras behavioristas norte-americanos partiam do pressuposto que nascemos sem qualquer conhecimento prévio, e que nosso aprendizado e desenvolvimento psicológico acontecem através da experiência de contato com o mundo exterior. Essa posição, que na filosofia é conhecida como *tábula rasa,* remonta a Aristóteles e suaontologia materialista, e foi difundida pela escola estoica até chegar nos empiristas ingleses da modernidade, onde ganhou força nas obras de John Locke e Thomas Hobbes ao contestar o inatismo cartesiano. Tal posição era também reflexo das ideias políticas defendidas pelos ingleses

da época: se todos os homens nascem sem qualquer forma ou conteúdo mentais, logo todos nascem iguais, e desta forma o poder não poderia ser algo divino (como defendiam os absolutistas).

Agora ficou mais fácil de compreender a posição de Turing: estando na Inglaterra, seu behaviorismo não era cego, pois ele compreendia que uma parte do processo de aprendizagem em humanos ocorre através da imitação, então se uma máquina for capaz de aprender,teria que ser também através da imitação. Qual a distância entre imitar um pensamento e pensar? Se nós humanos aprendemos imitando os outros humanos, porque um computador imitando humanos não pode também aprender? E se pode aprender, o que o impede de se tornar consciente? Para Turing não havia diferença entre imitar o pensamento e efetivamente pensar, pois pensamento humano e pensamento no geral não são também distintos, todos eles são adquiridos a partir da imitação e condicionamento comportamental.

> An important feature of a learning machine is that its teacher will often be verylargely ignorant of quite what is going on inside, although he may still be able tosome extent to predict his pupil's behavior. This should apply most strongly to thelater education of a machine arising from a child machine of well-tried design (orprogramme).(TURING, 1950, p. 460).

Turing se tornou tão lendário no universo da computação, que todos os anos o *Cambridge Center for Behavioral Studies of Massachusetts* oferece o *Prêmio Loebner*[7], que premiacom uma recompensa em dinheiro criadores de softwares capazes de enganar juízes humanos num teste de Turing. Alguns softwares já conseguiram enganar os juízes em perguntas simples, porém o prêmio principal oferecido para um software que for capaz de enganar um humano com uma resposta obtida através do processamento de imagem, texto e áudio ainda não foi ganho por nenhum competidor. Quando isso acontecer, o prêmio será oficialmente encerrado. Curiosidade o fato de que com bastante frequência os juízes confundem humanos com computadores no teste, dando peso pro argumento de que seres humanos não são critério muito bom para reconhecer outros seres humanos.

Nos anos 50 a tecnologia ainda não permitia a concepção de máquinas capazes de participar de um teste de Turing, de modo que seu idealizador

[7]http://www.eecs.harvard.edu/~shieber/Biblio/Papers/loebner-rev-html/loebner-rev-html.html

morreu sem poder ver na prática o teste aplicado. As teorias simbolistas ganharam mais força conforme as condições técnicas evoluíram no sentido do aumento do poder computacional, e o primeiro software com pretensões de inteligência artificial (*Logic Theorist*[8]) nasceu em 1956, apenas dois anos após a morte prematura de Turing, e poucos meses antes de John McCarthy cunhar o termo “Inteligência Artificial”, na Conferência de Dartmouth. “Artificial Intelligence is (...) find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves”. (MCCARTHY, 1955).

O paradigma *simbolista* se tornou tão forte que a própria definição de AI dependia dele. Dentre os cientistas criadores do *Logic Theorist* estavam Allen Newell e Herbert Simon, que publicaram nos anos 70 importantes artigos sobre a teoria simbolista, formando a base teórica para os futuros desenvolvimentos da computação digital.

2.2 – A posição simbolista de Newell e Simon

A criação do *Logic Theorist* representa uma nova posição filosófica no que diz respeito a teoria da mente, conhecida como *computacionalismo*. A prova empírica de que uma máquina pode realizar tarefas inteligentes, até então restritas aos seres humanos, era o indicador de que elas poderiam ter algum tipo de autonomia. Outro ponto chave dentro do estudo das teorias da AI é se essas inteligências deveriam ou não ser iguais a mente humana, ou seja, se inteligência de máquina é uma imitação da humana ou um modelo próprio baseado na estrutura “fisiológica” da mesma. Ao afirmar que um sistema material pode conter as propriedades necessárias de uma mente, se propuseram a resolver o problema corpo/mente. Essa posição mais tarde será chamada por Searle de “AI Forte”, e será tratada no devido momento deste trabalho.

Apesar da posição simbolista, Newell e Simon discordam do behaviorismo estrito de Turing, e apresentam uma visão inovadora dentro das

[8]https://history-computer.com/ModernComputer/Software/LogicTheorist.html

teorias da inteligência. Não existiria um princípio da inteligência, mas sim capacidades inteligentes, desta forma, a inteligência é compreendida como uma *propriedade emergente* de um sistema composto capaz de armazenar e manipular símbolos, colocando a atividade simbólica como centro da dela.

> Symbols lie at the root of intelligent action, which is, of course, the primary topic of artificial intelligence. For that matter, it is a primary question for all of computer science. [...] We measure the intelligence of a system by its ability to achieve stated ends in the face of variations, difficulties and complexities posed by the task environment.(NEWELL& SIMON, 1976, p.114).

Símbolos podem ser armazenados fisicamente, estruturados em formato de dados e processados em larga escala através de expressões relacionais, onde o papel da inteligência seria compará-los com descrições físicas fornecidas pelos sentidos. Um sistema simbólico físico[9] é então um sistema que contém dados e conjuntos de expressões aplicáveis a eles, de modo que a interação entre eles permitiria a existência da inteligência como propriedade emergente de tal sistema. Para Newell e Simon a inteligência pode surgir de qualquer sistema simbólico físico que tenha sido projetado para tal finalidade, e apontam humanos e computadores como bons exemplos de tais sistemas. A seguinte frase é famosa e ilustra bem a posição: “A physical symbol system has the necessary and sufficient means for general intelligent action”, (NEWELL & SIMON, 1976, p.116).

2.3 – John Searle e o Naturalismo Biológico

Um dos mais importantes e conhecidos críticos da *posição simbolista* foi o filósofo John Searle, que buscou contra argumentar os pressupostos behavioristas da teoria através do experimento mental da sala chinesa, servindo de base para sua posição que ficou conhecida como *Naturalismo Biológico*, que levou pro centro da discussão a importância e necessidade dos substratos biológicos para a ocorrência da inteligência.

Searle foi um filósofo de matriz analítica, e que começou seus estudos pela filosofia da linguagem a partir da obra de Frege. A principal questão da

[9]http://ai.stanford.edu/users/nilsson/OnlinePubs-Nils/PublishedPapers/pssh.pdf

filosofia da linguagem deveria ser como o conjunto de sons que nós produzimos se relaciona com a realidade. Em outras palavras: de onde vem o significado que nós atribuímos às palavras? Mesmo sendo a linguagem centro do nosso conhecimento e relação com a realidade, são as próprias palavras também fundadas em bases orgânico-biológicas. Desta forma seria necessário diferenciar meras *palavras* de *atos de fala*.

Sua entrada no campo da filosofia da mente foi quase acidental. Tentando descobrir como as ondas acústicas se transformam em atos cognitivos (como da física a linguagem chega até a semântica), acabou se deparando com o problema mente/corpo e com as respostas clássicas da filosofia. Tal direcionamento filosófico nos evidencia o fato de que o *Naturalismo Biológico* é uma posição monista sobre a relação mente/corpo. Parte do pressuposto que não existe separação entre ambos e que estados mentais correspondem a estados físicos. Rejeitando as posições clássicas que vão desde o materialismo estrito até os dualismos mais metafísicos, assenta suas bases na neurociência e na biologia. O que temos fisicamente é um conjunto de células que trabalham em rede, e do trabalho de suas reações neuroquímicas emergiria uma propriedade subjetiva a qual chamamos *mente*.

> (...) podemos resumir esta questão dizendo que a consciência não é redutível da maneira que outros fenômenos são redutíveis, não porque o modelo de fatos no mundo real envolva algo de especial, mas porque a redução de outros fenômenos depende em parte da distinção entre “realidade física objetiva”, de um lado, e meras “aparências subjetivas”, de outro; e da eliminação da aparência dos fenômenos que foram reduzidos. Mas no caso da consciência, sua realidade é a aparência. (Searle 1992/2006, p. 176)

A principal crítica de Searle ao simbolismo de Turing diz respeito a seu principal pressuposto: não discorda do fato do cérebro ser uma máquina, mas não concorda que uma máquina que apresente comportamento idêntico ao humano esteja efetivamente pensando, e para exemplificar sua argumentação recorre a outro experimento (facilmente realizável com a tecnologia que dispomos hoje) chamado de *experimento da sala chinesa*.

Neste experimento mental, Searle estaria dentro de uma caixa onde receberia por uma abertura textos em chinês, que seriam comparados com tabelas de símbolos que ele possui mas não compreende o significado. Apenas

seguindo as instruções em inglês que acompanham os símbolos, ele seria capaz de formular respostas em chinês e devolver por outra abertura da caixa, mesmo sem entender uma palavra de chinês. Para quem estivesse observando externamente a caixa, pareceria que ela está compreendendo chinês e respondendo, mas na verdade ela estaria apenas fazendo uma comparação de símbolos sem compreensão de seu significado. Devido a isso, afirma que não podemos aceitar que a capacidade simbólica seja suficiente para ocorrência de inteligência, no máximo que seja uma propriedade necessária, mas não suficiente, e evidência disso seria o fato da caixa estar manipulando símbolos cujo significado desconhece.

Mesmo que o experimento de Turing apresente o resultado esperado, não prova que as máquinas estão se comunicando conscientemente, pois estão apenas aplicando regras formais as quais desconhecem o conteúdo.Tal atividade não corresponderia a pensamento, pois inteligência depende de habilidades *sintáticas*(formas) e *semânticas*(conteúdos):

> A razão por que nenhum programa decomputador pode alguma vez ser uma mente é simplesmente porque um programa decomputador é apenas sintático, e as mentessão mais do que sintáticas. As mentes sãosemânticas, no sentido de que possuemmais do que uma estrutura formal, têm umconteúdo. (SEARLE, 1997, p. 38-9).

Posto isso, a pergunta agora se modifica para: qual a origem das habilidades semânticas da inteligência? Para responder se faz necessário recorrer a ideia de Intencionalidade.Conceito filosófico bastante famoso, sua origem é medieval mas quem o tornou popular foi Franz Brentano, ao utilizá-lo para designar uma propriedade que é única e exclusiva das mentes: o fato de que apenas os fenômenos mentais representam (significam) algo. Através da Intencionalidade nossa mente representa objetos que lhe são fornecidos pelos dados dos sentidos, e atribui valores de significado a eles, de modo que nossa habilidade semântica depende diretamente dela.

A Intencionalidade também coloca Searle contra a posição de Newell e Simon, pois para eletal faculdade depende não apenas de cruzamento de dados (desta forma o Google Translate já teria se tornado consciente), mas também de crenças e desejos relacionados ao conteúdo apreendido, e crenças

e desejos envolvem *conteúdos* mentais. Uma máquina puramente formal iria apenas substituir um símbolo por outro, sem jamais criar alguma crença em relação a eles.

Cabe citar que também existe uma discussão sobre *Intencionalidade Original* e *Intencionalidade Derivada* que envolve Searle e Dennet: para Dennet a intencionalidade das mentes humanas não seria *Original* como imaginava Searle, mas derivada. Desta forma, se a intencionalidade da mente humana é derivada, uma inteligência de máquina puramente formal poderia de alguma forma apresentar intencionalidade. Porém para que isso aconteça seria necessário criar uma máquina com capacidades sensíveis próximas das humanas, e esta máquina teria que aprender com o ambiente também e ser treinada para tal finalidade. Dennet até imagina uma máquina com sensores visuais, braços e sentidos artificiais análogos aos humanos, e afirma que talvez essa seja a melhor forma de criarmos uma inteligência artificial que satisfaça os critérios da posição naturalista biológica.

Neste ponto existe uma certa concordância com Turing, que afirmava que o provável melhor caminho para a criação da AI seja através de máquinas capazes de aprender. Os algoritmos de *machine learning* comuns hoje em dia estão nos mostrando que os melhores resultados práticos em tarefas de AI acontecem quando combinamosmodelos lineares puramente simbolistas com algoritmos relacionais de redes de dados. Tais algoritmos são um pouco mais recentes, e nasceram de uma corrente de pensamento conhecida como *Conexionismo*.

2.4 – Posição Conexionista

No que diz respeito a teoria da computação, o *Conexionismo* é o paradigma científico hegemônico dos nossos tempos. Suas bases teóricas remontam a metade do século passado, mas conquistou evidência somente na última década devido as condições tecnológicas permitirem sua utilização em larga escala somente nos tempos recentes. Por necessitar de uma grande quantidade de informações para treinamento e atribuição de pesos aos dados, tais algoritmos exigem um grande poder de processamento, algo que só se tornou tecnologicamente viável nas últimas duas décadas. Somadas ao fato de

grande parte do poder econômico mundial estar concentrado em empresas de tecnologia que dispõe de bancos de dados gigantes de clientes (tais como as redes sociais), temos a receita de como o *Conexionismo* ganhou tal protagonismo: reconhecimento de voz, reconhecimento facial, leitura biométrica e tantos outros recursos tecnológicos que hoje tomam conta da paisagem urbana só foram possíveis graças aos avanços de tal paradigma. O exemplo dos carros de piloto autônomo é um caso *Conexionista* de uma tecnologia bastante recente, que possui a particularidade de ter sido treinado num ambiente também virtual.

O *Conexionismo*tem suas bases no trabalho de Warren McCulloch e Walter Pitts da *University of Chicago*. Em 1943 eles provaram que sinais elétricos dos nossos neurônios podem ser modelados no formato de expressões lógicas[10]. Partindo de tal evidência, em 1958 o psicólogo Frank Rosenbratt criou um algoritmo que buscava reproduzir a estrutura das redes neurais de nosso cérebro, tentando simular asdiversas camadas de processamento de informações. Assim nasceu o *Perceptron*[11], um algoritmo que no lugar de símbolos representacionais, utiliza cargas elétricas e seus respectivos pesos distribuídos em camadas para representar dados. Diferentemente dos algoritmos lineares simbolistas, agora os dados poderiam ser salvos de forma distribuída, e processados mesmo que incompletos ou inconsistentes, de forma paralela e por mais de uma unidade de processamento ao mesmo tempo.

A partir de então teve início a chamada *Revolução Conexionista*, que adentrou as décadas seguintes com a publicação das obras de MacClelland e Rumelhart, com a ideia básica de criar uma grande rede de unidades computacionais simples (*perceptrons*) onde ocorreria processamento de dados, tal qual ocorre com os neurônios em nosso cérebro. Enquanto o *paradigma simbolista* era baseado na representação, o *conexionismo*tinha suas bases fundadas no processo biológico de aprendizagem, dependendo de grandes quantidade de dados para analisar e aprender seus padrões, que ficam armazenados de forma distribuída nas unidades de processamento da rede.

O fácil acesso a grandes quantidades de dados que as redes sociais e

[10]https://link.springer.com/article/10.1007/BF02478259
[11]https://psycnet.apa.org/record/1959-09865-001

serviços digitais criaram em nossa época levaram tais modelos a seu apogeu prático, criando áreas de trabalho sobre análise e inteligência de dados, o chamado *deep learning*. Tais algoritmos fazem reconhecimento de imagens baseados no modelo funcional de nossos olhos, processam sons baseados em modelos digitais que simulam nossos ouvidos, e tomam decisões baseados em pesos atribuídos pelo processamento de seus neurônios artificiais, funcionando de forma análoga a nossos neurônios e sinapses.

Talvez o ponto conceitual mais fraco de tal posiçãoseja não ter respondido a principal crítica de Searle.O problema fundamental das inteligências de máquina que é inerente a todos os paradigmas é o mesmo: as máquinas operam apenas no nível sintático e não são capazes de chegar aosemântico. Mesmo tendo seu modelo inspirado na estrutura do cérebro, as redes neurais artificiais acabam por abstrair partes muito importantes da nossa cognição, tais como a influência biológico-hormonal e a neurotransmissão. Diferentemente dos *Perceptrons*, nossos neurônios não são homogêneos e reagem de formas diferentes ao mesmo conjunto de cargas elétricas que outros neurônios de outra parte de nosso cérebro, assim como várias propriedades das redes artificiais também não são encontradas nas nossas mentes.

Desta forma talvez seja mais seguro afirmar que as redes neurais artificiais copiam apenas partes de nossa atividade cognitiva, simulando somente parcelas da nossa inteligência. Por isso se faz tão necessária a fusão de modelos e utilização de paradigmas híbridos: a nossa própria cognição opera fazendo uso de ambos. Quem sabe quantos outros mecanismos ainda não descobertos os próximos anos nos trarão no campo da cognição humana? Parece que a AI se encontra num ponto análogo ao da Psicologia do início do século passado, onde vários paradigmas competiam entre si, e no final a melhor abordagem se mostrou híbrida e adaptada a cada caso específico. Talvez ainda falte um pouco de amadurecimento ao campo dado o quanto é recente sua emergência, mas os próximos capítulos desta epopeia se mostram muito promissores.

2.5 – As Críticas ao Representacionalismo

Nos últimos tópicos busquei falar de forma resumida da história da AI e de seus pressupostos epistemológicos, através de momentos-chave onde ocorreram avanços significativos. Propositalmente não falei dos momentos onde a AI encontrou-se estagnada (também chamados de*AI Winter* – Inverno das AI), pois compreendo que tais momentos precisam ser analisados a partir do viés crítico. Nesta passagem entre o otimismo científico que chegou nos avanços aqui mencionados até o pessimismo Heideggeriano tratado na primeira parte do texto, veremos que os critérios de caracterização do estudo das AI passa justamente por momentos de otimismo seguidos de estagnação. Os anos de *AI Winter* estão diretamente relacionados a mudança da *essência da verdade* e a *entificação do ser*, que tornaram a representacionalidade paradigma hegemônico do meio científico e das ciências no geral.

Tivemos anos de grandes avanços seguidos de anos de grandes estagnações nos conhecimento das AI, e esses momentos de estagnação são conhecidos como *AI Winter*[12]. O termo foi utilizado pela primeira vez em 1984 na conferência AAAI (Association for the Advancement of Artificial Intelligence). Sua definição estrita ficou estabelecida como “Um período de declínio de entusiasmo após tempos de sucesso no campo da AI”. Desta forma, a evolução da área enfrentou alguns períodos de AI Winter, documentados como períodos de 1973 a 1979 e de 1988 a 1994. Discute-se a possibilidade de estarmos em novo período de crise no campo, e existe bastante material sendo produzido a nível internacional.

A crítica de Heidegger é apoiada na tese de que a metafísica ocidental criou uma ciência que se importa apenas com os*fenômenos vulgares*, e atinge a visão de homem que a fundamenta, e por consequência também a Psicologia e os modelos que serviram de fundamentação aos teóricos das *Inteligências de Máquina*. Em uma das conferências dos *Seminários de Zollikon*, aponta através de exemplos práticos como a visão mecanicista, herdada da modernidade que busca reduzir o ser humano a uma máquina de estímulo resposta, acabou tomando conta das ciências:

[12]https://ai.stanford.edu/~nilsson/QAI/qai.pdf#section.24.4

> O exato das ciências exatas não pode ser exatamente determinado, ou seja, por vias de cálculo, mas apenas ontologicamente, e assim sendo o tipo de verdade que se aplica à ciência é no sentido da ciência natural exata. Sua verdade é verificada pela eficiência de seus resultados. Quando este modo de pensar determina o conceito científico de ser humano e o investiga a partir do modelo de conjunto de leis, como agora acontece na cibernética, em pouco tempo a destruição do ser humano será perfeita. É por isso que eu combato a ciência, mas não a ciência enquanto ciência, mas apenas o caráter absolutizante da ciência natural. (HEIDEGGER, 2009, p. 197).

AI Simbólica é muito boa para resolver problemas que envolvem abstração e comparação de dados. Conexionismo se mostra mais promissor ao tentar resolver problemas que exigem interação com grandes massas desordenadas de dados. A maior parte dos problemas do mundo real se encontra distribuída em algum ponto entre os dois paradigmas, e modelos híbridos estão ganhando espaço nas pesquisas e softwares comerciais, mostrando que na verdade os paradigmas não se encontram em oposição, mas que precisam trabalhar de forma complementar e nas tarefas onde cada um deles se mostra mais eficiente.

Tais algoritmos são pré-concebidos para finalidades específicas, e se mostram muito superiores a racionalidade humana em tais tarefas, tornando-se mais eficientes e baratos. Tais propriedades (eficiência e baixo custo operacional) despertam o interesse dos grandes blocos econômicos, que através de grandes investimentos em desenvolvimento tecnológico se empenham em levar ao mundo soluções que sejam capazes de automatizar tarefas que talvez nem humanos com grande saber técnico poderiam realizar.

A natureza de tais artefatos os torna complexos conjuntos de várias áreas do saber, que vão desde projetistas, passando por desenvolvedores de software, equipes de teste e adaptação de uso, até o momento de seu contato com o mundo real. Tal multiplicidade tornou necessária sua análise conceitualpara alcançar uma compreensão funcional que possa se conectar a nossos conceitos filosóficos, tornando possível compreender seus níveis de autonomia, dependência e determinismo ao ponto de elucidar a natureza causal de suas ações nas sessões seguintes do trabalho.

Todas linhas de pensamento estudadas até agora partem do mesmo pressuposto: a mente é criada pela matéria, portanto são de alguma forma *materialistas* e *reducionistas*. Daí nasce a aspiração de simular a máquina-cérebro em uma máquina digital criada pelo homem, e mesmo a tarefa tendo se mostrado muito mais árdua do que o imaginado, progressos significativos ocorreram no último século.

Tal *materialismo* e *reducionismo* são a base de toda nossa ciência instrumental, e tal afirmação encontra-se em sintonia com as ideias Hedeggerianas da *entifição do ser* e da mudança na *essência da verdade* através da elevação dos *fenômenos vulgares*. Como vimos na primeira parte do capítulo, ambas ideias convergem no caminho que o pensamento da civilização ocidental seguiu desde os pré-socráticos, passando por Platão e Aristóteles, somado aos avanços que a ciência teve nos dois grandes saltos dos séculos XVIII e XX. Tal caminho do *esquecimento do ser* que nos converteu em civilização técnica fez com que a representação tomasse conta das ciências no geral, e talvez isso explique porque nossas inteligências artificiais ainda não apresentam qualquer sinal de consciência: toda nossa tecnologia é representacional e formal, enquanto a consciência é semântica.

> O ser não é pensado em sua essência desveladora, isto é, em sua verdade. Entretanto, a metafísica fala da inadvertida revelação do ser quando responde a suas perguntas pelo ente enquanto tal. A verdade do ser pode chamar-se, por isso, o chão no qual a metafísica, como raiz da árvore da filosofia, se apóia e do qual retira seu alimento. Pelo fato de a metafísica interrogaro ente, enquanto ente, permanece ela junto ao ente e não se volta para o ser enquanto ser (HEIDEGGER, 1996, p. 77-78).

Além da crítica semântica, Searle também nos deu fortes argumentos para acreditar que uma mente artificial exige uma estrutura física que a comporte e produza, já que a mesma é uma propriedade emergente de um sistema complexo. As inteligências de máquina derivadas do paradigma simbolista são totalmente lineares, e justamente por isso se mostram pouco eficientes no que diz respeito a manipulação de dados do mundo real, sendo mais assertivas em atividades matemáticas e lógicas. Das suas deficiências surgiu a necessidade de uma nova forma de conceber AI: o conexionismo. Essa nova forma nasceria dos avanços da medicina e da neurociência, levando

alguns cientistas a aspirar a criação de uma máquina capaz de simular nossa estrutura e atividades neurais. Porém o *Perceptron* se mostrou incapaz de dar conta de toda complexidade envolvida em criar uma consciência artificial, e talvez o principal motivo seja fato de a ciência ainda não ter evoluído algumas áreas necessárias o suficiente para a correta compreensão da emergência da consciência.

Colocadas as IA dentro de suas matrizes epistemológicas, podemos afirmar com certa propriedade que o modo de compreender suas ações não pode ser o mesmo modo que usamos para compreender as ações humanas, pois sua inteligência é muito diferente da nossa (linear, enquanto a nossa é semântica).

Conscientes ou não, a realidade em que vivemos torna inevitável que tais máquinas, ao executar tarefas complexas a nível social tenham que, em algum momento, tomar uma decisão sobre alguma questão moral. Quando chega esse momento decisivo entra a pergunta: as máquinas são capazes de aprender os conceitos de certo e errado? Ou melhor dizendo, agentes artificiais podem ser agentes morais? Alguns estudiosos da área da cognição afirmam que não demorará muito para que tenhamos que começar a compreender eticamente tais máquinas.

3 A MORALIDADE NOS AGENTES ARTIFICIAIS

Anteriormente levantamos hipóteses apoioadas nas teorias de Heidegger, que nos ajudama entender porque esses quase 70 anos de existência das AI foram tão árduos: suas limitações encontram-se num nível ontológico mais abstrato do que seus próprios paradigmas.Todas linhas de pensamento convergem para um *materialismo reducionista*, justamente por suas fundamentações estarem nos *fenômenos vulgares* da ciência instrumental.

Também ficou claro que as limitações das *Inteligências de Máquina* são gerais e inerentes a todos os paradigmas, pois até agora não foi possível criar uma inteligência artificial realmente consciente, capaz de intencionalidade semântica e que passe pelas críticas de Searle. Porém isso não significa que as máquinas, em certa medida e em determinadas situações, não possam ser consideradas agentes morais, pois é inevitável o fato de que suas ações no mundo esbarrem hora mais hora menos em decisões de cunho moral.

Como vimos no início deste trabalho, a tradição filosófica, na maior parte dos casos, associa responsabilidade moral ao conceito de consciência. Isso ocorre devido aos fatores citados por Heidegger (*hegemonia científica da técnica*, *esquecimento do ser* e *verdade como adequação*). Após analisar os tipos de inteligências artificiais existentes, concluímos que até o momento nenhum deles atingiu status de consciência semântica intencional. Desta forma, compreendendo moralidade dentro do discurso da tradição, as máquinas se tornam inimputáveis de tal responsabilidade. Então quais os critérios necessários (se é que existem) para que um agente artificial, livre dos conceitos da tradição, possa ter suas ações compreendidas dentro da esfera da moralidade?

No artigo *Moral machines: Teaching robots right fromwrong*, Wallach e Allen(2009)propõem que num futuro, onde as inteligências de máquina tenham se aperfeiçoado ao ponto de tomar decisões de natureza complexa e independente da supervisão humana, seja o momento adequado para incluir tais máquinas dentro do que compreendemos como agentes morais. Para conceituar tais casos, criam o conceito de Agente Moral Artificial – AMA. Como veremos adiante, o rápido avanço da computação acabou por transformar

talconceito para muito além da ideia de seus criadores.

3.1 – A Questão da Linguagem

No experimento de Turing, capacidades linguístico-comunicativas (possíveis indicadores de intencionalidade) são demarcadores que diferenciam inteligência de não inteligência. Porém tais capacidades são compreendidas apenas a partir do caráter *apofântico* da linguagem (pois neste caso tal capacidade seria mero produto da imitação), ignorando seu caráter *hermenêutico*, a partir do qual o Dasein atribuisignificado aos fenômenos, constituindo seu horizonte de sentido. Em outras palavras, a significância é a estrutura do mundo, que acontece dentro de uma rede referencial acessada pela manualidade, que vai muito além da mera sintaxe da linguagem a qual as máquinas possuem acesso. A interpretação (*auslegung*) acontece a partir da união das dimensões apofântica e hermenêutica do lógos, abrindo o Dasein ao seu horizonte de compreensão.

> O caminho para a linguagem - isto soa como se a linguagem estivesse bem longe de nós, em alguma parte, de modo que para lá chegar ainda teríamos que nos pôr a caminho. Será mesmo necessário um caminho para a linguagem? Segundo uma antiga tradição, nós somos aqueles seres capazes de falar e, assim, aqueles que já possuem a linguagem. A capacidade de falar, ademais, não é apenas uma faculdade humana, dentre muitas outras. A capacidade de falar distingue e marca o homem como homem. Essa insígnia contém o desígnio de sua essência. O ser humano não seria humano se lhe fosse recusado falar incessantemente e por toda parte, variadamente e a cada vez, no modo de um "isso é", na maior parte das vezes, impronunciado. À medida que a linguagem concede esse sustento, a essência do homem repousa na linguagem. (HEIDEGGER, 2003b, p. 191).

Mais do que apenas vinculada às proposições lógico-gramaticais, a linguagem possui um horizonte hermenêutico, no qual nós já estamos sempre pré-inseridos, e que por isso constitui estrutura à priori da mesma. Tal postura implica uma mudança de caráter de *Antropologia Filósófica*,pois significa

compreender homem não mais como uma coisa carregada de propriedades, e sim como um acontecimento interpretativo, inserido num horizonte histórico de possibilidades que se apresenta como mundo possível, acessado e compreendido através da manualidade.

Ao buscar elucidar como se constitui a mundanidade (*Weltlichkeit*), Heidegger argumenta que o mundo não nos é dado previamente como uma série de coisas dotadas de propriedades, mas antes como uma série de entes com os quais o *Dasein* se relaciona e atribui significados, inseridos numa totalidade de sentido que se encontra a todo momento dotada de disponibilidade (*Zuhandenheit*). Para além da linguagem, a significância se apresenta como estrutura formal da mundanidade do mundo, e totalmente inacessível às inteligências de máquina, pois a mesma ocorre dentro de uma estrutura intencional, exigindo como pré-requisito que o agente possua intencionalidade.

> O mundo mais próximo [do ser-aí cotidiano] é o mundo circundante (Umwelt). Para se chegar à idéia de mundaneidade [em geral], a investigação seguirá o caminho que parte desse caráter existencial do ser-no-mundo mediano. Passando por uma interpretação ontológica dos entes que vêm ao encontro dentro do mundo circundante, poderemos buscar a mundaneidade do mundo circundante (HEIDEGGER, 1989, p. 114).

Tal disponibilidade dá-se sempre dentro de um contexto, dentro do qual os entes se tornam adequados às circunstâncias conformativas. A adequação conformativa constitui uma rede referencial a partir da qual a significância se forma e reforça, ou seja, os entes se apresentam já desde sempre carregados de significado, inseridos num horizonte de sentido sempre disponível ao Dasein, e a partir do qual se fundamenta a possiblidade de sua linguagem. Mesmo livrando as máquinas da necessidade de consciência e de inteligência como a humana, ainda não somos capazes de compreedê-las moralmente através do pensamento de Heidegger, pois o Dasein é único e singular.

Aplicando seu método de *Destruktion*, tenta remontar a linguagem de forma fenomenológica, ou seja, livre dos preconceitos da tradição, de modo que sua base ontológico-existencial seja desvelada. Sempre voltando aos gregos, recupera a noção de λόγοςpredominante entre eles, que significaria

originariamente relação com o ser expressada de forma linguística, ou seja, seu caráter primário é de origem compreensiva, anterior a sua estrutura lógico-semântica. Assim sendo, a linguagem possui uma dupla estrutura, chamada de estrutura *hermenêutico-apofântica*. Isso implica que antes de sermos capazes de expressar algo em forma de signo linguístico, de alguma forma nós já tivemos acesso ao ente correlato e absorvemos informações provenientes de sua rede referencial, onde seu significado encontra-se codificado pelo uso.

> A ontologia antiga tem por solo exemplar de sua interpretação de ser o ente que se encontra dentro do mundo [isto é, intramundano]. Como modo de acesso a ele vale o νοεῖνou entãoολόγος. É aí queela encontra o ente. O ser desse ente precisa, porém, ser apreendido em um λέγειν (deixar ver) distinto, de modo que esse ser seja compreensível antecipadamente como aquilo que ele é e em cada ente já é. A interpelação já sempre prévia do ser na discussão (λόγος) do ente é o κατηγορεῖσθαι (HEIDEGGER, 1989, p. 44).

Tal afirmação coloca o principal empecilho na possibilidade das inteligências de máquina compreender os signos linguísticos: o significado não está totalmente embutido nas palavras, mas ao menos uma parte dele é formada quando o Dasein acessa a rede referencial, pois neste momento ele une a significância que encontra na estrutura da mudanidade do mundo a todo campo de significado que habita sua mente. Afirmar que isso impossibilita totalmente a moralidade nas inteligências de máquina me parece bastante precipitado, todavia nos obriga a pensar ao menos os limites de tais inteligências, dado que ao menos uma parte da rede de significado talvez precise ser pré-embutida nelas (justamente um dos modelos de AMAs que veremos a frente constitui-se desta maneira).

Provavelmente nossas máquinas e nosso saber ainda não foram capazes de se libertar das limitações empregadas pelo modo como nossa civilização se fundou e evoluiu nos últimos séculos, de modo que é possível que estejamos esgotando os limites do paradigma representacional herdado do *Iluminismo*. Quando nossa ciência encontrar caminhos para além da representação técnica, talvez se abra todo um novo campo de possibilidades para o aperfeiçoamento das AI no geral.

É na τέχνηé que a *verdade* (άλήθεια)[13] se apresenta, e por isso tal saber, com o passar dos séculos e com a evolução dos métodos empíricos, acabou por influenciar fortemente a concepção clássica de *ciência*, nos levando a uma visão onde essa correspondência de *verdade* com *técnica* cristalizou sua dependência de avanços tecnológicos à representação, limitando seu horizonte de compreensão a somente aquilo que se valida nos moldes do *ente*. Para Heidegger, a consciência através da linguagem (λόγος discursivo) é característica única e exclusiva do Dasein, constituindo sua essência como Ser, logo nenhuma máquina seria capaz de simular tal faculdade, menos ainda uma máquina linear baseada em representação e sem propriedades semânticas, que jamais vai ser capaz de consciência e questionamento de sua própria existência, pelo simples fato de não poder interpretar semanticamente a linguagem.

Olhando para a história da Ciência da Computação, nós da Filosofia podemos ter certa surpresa ao descobrir que o assunto da moralidade de agentes artificiais não é novidade. O termo AMA (Agente Moral Artificial – Artificial Moral Agent) foi cunhado em 2009[14],e desde então se tornou um conceito chave dentro do campo da Filosofia conhecido como Machine Ethics. Na época a principal preocupação de seus autores era atentar pro fato de que a evolução das tecnologias chegaria um dia no ponto onde seria necessário pensar uma moralidade para agentes artificiais. Os acontecimentos da década seguinte tornaram tal visão real: a larga disponibilidade de dados que as redes sociais e plataformas da internet possuem aceleraram o processo de aperfeiçoamento das máquinas ao ponto de hoje ser real e rotineiraa convivência com inteligências de máquina em nosso dia a dia.

Em 2009 não era possível prever a dimensão e velocidade que o avanço das tecnologias da informação tomariam, de modo que o futuro previsto está chegando cava vez mais rapidamente. Tal velocidade nos obriga a rever o entendimento sobre os AMAs frente os novos avanços da informática, tanto em

[13]A técnica não é, portanto, um simples meio. A técnica é uma forma de desencobrimento. Levando isso em conta, abre-se diante de nós todo um outro âmbito para a essência da técnica. Trata-se do âmbito do desencobrimento, isto é, da verdade. (...) Técnica é uma forma de desencobrimento. A técnica vige e vigora no âmbito onde se dá descobrimento e desencobrimento, onde acontece άλήθεια, verdade. (HEIDEGGER, 2002, pp.16 – 17).

[14] Para mais informações: AAAI Fall Symposium (https://www.aaai.org/Library/Symposia/Fall/fs05-06.php

seu status moral como em seu status ontológico. Se os *Agentes Artificiais* baseados em *machine learning*não possuem o mesmo status ontológico que o ser humano, e também não possuem o mesmo que uma máquina linear, então qual é o status ontológico de tais entes?

3.2 – AMA: Agente Moral Artificial

O conceito de Agente Moral Artificial, como dito anteriormente, foi cunhado por Colin Allen e Wendell Wallachno *AAAI Fall Symposium*, em 2009. Sua criação implicou uma clara divisão dentro dos paradigmas da AI. O debate esteve sempre centrado na discussão entre as duas visões hegemônicas, chamadas de Visão Convencional (*standard view*) e Visão Funcionalista (*functionalist view*). A visão convencional estaria alinhadacom as teoriasde Searle e a tradição filosófica, que inferem que um agente deva ter racionalidade, intencionalidade e autonomia para pode ser considerado moral. Enquanto a visão funcionalista ficaria mais próxima do modelo de Turing, onde seria o comportamento indicador suficiente de capacidades de agência e moralidade.

Para Wallach e Allen (2009), mais importante que a discussão do status ontológico, é a criação de modelos éticos capazes de abarcar os diversos tipos de agentes artificiais existentes, ou que por ventura possam vir a ser criados. Argumentam que chegamos num ponto de avanço tecnológico onde se tornou socialmente necessário implementar modelos éticos para as ações dos sistemas autônomos, independente de todas discussões teóricas a cerca do tema. Atualmente todos sistemas de AI existentes são vazios no que diz respeito a moralidade (*ethically blind*), pois sua tomada de decisão não envolve qualquer tipo de camada moral. Apontam como necessário criar regras éticas que possamos embutir na programação das máquinas, pois a crescente evolução e aperfeiçoamento de tais equipamentos está criando situações cada vez mais críticas, onde ações de entidades autônomas se vêem perante decisões morais.

> It is easy to argue from a position of ignorance that the goal of artifcial moral agency is impossible to archieve. But precisely what

> are the challenges and obstacles for implementing artificial morality? There is a need for serious discussion of this question. The computer revolution is continuing to promote reliance on automation, and autonomous systems are ingreasingly in charge of a variety of decisions that have ethical ramifications. How comfortable should one be about placing one's life and well-being in the hands of ethically ignorat systems?(WALLACH & ALLEN, 2009, p. 14-15).

Já a visão funcionalista, em seus desdobramentos mais recentes, se faz voz na argumentação de Floridi e Sanders (2004). Utilizam o modelo comportamental justamente para fugir dos critérios convencionais de consciência e intencionalidade, buscando criar uma *mind-less morality* (Floridi e Sanders,2004 p. 351), ou seja, uma moralidade que não dependa de uma mente, seja ela humana ou não. Nesta visão, o nível de abstração das interações dos entes com seu meio serve como critério geral para delimitar os paradigmas de agência moral. Os autores propõem o seguinte modelo para a agência moral:

a) O ente deve ser capaz de interagir com seu ambiente;
b) O ente deve ter a capacidade de mudar a si mesmo a suas interações de forma independente de influências externas;
c) O ente pode se adaptar de acordo as informações coletadas de suas interações com o ambiente.

Dentro de tal delimitação, os sistemas de Machine Learning poderiam muito bem ser entendidos como Agentes Morais, pois obedecem a todos critérios. São capazes de interagir com o ambiente, e se adaptam e aprendem através da tentativa e erro. A base conceitual deste tipo de máquina é justamente sua suposta capacidade de aprendizado. Porém cabe discutir o ponto da independência, pois não existe nenhuma máquina 100% livre das pré-determinações de seu projetista.

Os modelos de AMAs que a ciência da computação tentou criar invariavelmente esbarram em uma das três grandes linhas de estudo da Ética. Quando removemos o fator social, mesmo nos casos de *machine learning*, a moralidade acaba parecendo apenas um conjunto de regras pré-estabelecidas, e olhando dessa perspectiva fica muito mais fácil imaginar uma máquina moral, afinal ela apenas precisaria seguir linear e cegamente as instruções de seus

algoritmos, funcionando como exemplo perfeito para os defensores das éticas normativas. Na literatura filosófica atéencontramos casos de tentativas de criação de bots baseados no utilitarismo[15], e também estudos teóricos de como seria um robô moral sem preceitos normativos (uma desejável *moralidade emergente*).

O caminho teórico a ser tomado pelo projetista de uma AI depende de muitos pressupostos que na maior parte dos casos ficam implícitos, pois perpassam campos muito amplos da Filosofia e da Ciência e vão criando seu corpo teórico movimentando-se entre epistemologia, ontologia, ética, processamento de dados e teoria da computação. Tal variedade de conhecimentos de ponta acaba por obrigar o profissional de pesquisas na computação a fazer invariáveis cortes. Vezes motivado por um bom estudo teórico, vezes motivado por convicções pessoais. Até que ponto um algoritmo é realmente imparcial?

Mesmo tendo reconstruído o caminho teórico que nos levou até o desenvolvimento dos AMAs, e tendo tornado mais evidentes seus pressupostos ontológicos, ainda é muito complicado delimitar exatamente em que ponto Agentes Artificiais se tornam morais. Porém ficou claro que dentre os agentes artificiais existem diferenças suficientes para abrir margem para que alguns deles se encontrem em situações de tomada de decisão moral. Nestes casos o agente artificial não pode ser eticamente cego, precisando ter alguma forma de acesso a algum tipo de moralidade, que torna suas ações mais próximas do código moral vigente. Para que isso seja possível, é necessário que o agente moral tenha sido projetado para tal finalidade, e ao longo da história recente da computação tivemos algumas tentativas que valem ser citadas.

3.3 – Machine Ethics e os tipos de AMAs

Tendo em mente que os AMAs possuem diferenças de nível ontológico entre si, fica claro que tais diferenças acabam por pré-definir as delimitações

[15]https://www.aaai.org/Library/Symposia/Fall/2005/fs05-06-006.php

técnicas que tornarão possível a tal ente desempenhar sua função prática. Desta forma, podemos afirmar que existem basicamente dois paradigmas técnicos a partir dos quais os atuais *Agentes Artificiais* hoje existem. É importante compreender estes paradigmas, pois eles demarcam os limites da suposta moralidade que será colocada a disposição de tal agente.

O primeiro paradigma técnico parte do pressuposto de que teríamos uma moralidade pré-embutida à priori, como um conjunto de instruções morais que devem ser seguidas. Tal paradigma cria *Agentes Artificiais* que possuem seu código moral pré-embutido de instruções sobre como agir perante as situações. Este seria um agente moral que não aprende, a desta forma seus programadores precisariam prever todas possíveis situações onde o mesmo se envolveria. Tal paradigma encontra-se diretamente relacionado ao *simbolismo*, pensando nesse caso numa máquina com comportamento ético.

Já o segundo paradigma está mais relacionado ao atual *conexionismo*, onde o agente aprende no processo. Neste caso o mesmo não possui um código moral pré-embutido, mas é programado de modo que seja capaz de aprender perante as situações que exijam tomada ética de decisão. Através do *machine learning* a máquina aprenderia com as situações de sua aplicação. Precisaria ser treinada em ambiente de testes até aprender uma quantidade segura de modelos de tomada de decisão que acabem por resultar o mais próximo da sua moralidade pré-suposta, para só então estar prontas para agir no mundo.

> A good moral agent is one that can detect the possbility of harm or neglect of duty, and can take steps to avoid or minimize such undesirable outcomes. There are two routes to accomplishing this: First, the programmer may be able to anticipate the possible courses of action and provide rules that lead to the desired outcome in the range of circumstances in which the AMA is to be deployed. Alternatively, the programmer might build a more open-ended system that gathers information, attempts to predict the consequences of its actions, and customizes a response to the challenge. Such a system may even have the potential to surprise is programmers with apparently novel or creative solutions to ethical challenges.(WALLACH & ALLEN, 2009, p. 16).

O campo de estudo conhecido como *Machine Ethics* possui também seus eixos comuns de estudo. O pressuposto básico é buscar formas de embutir ou de determinar comportamentos ou conceitos éticos em um ente com inteligência artificial. Vimos que as diversas formas de *Inteligências de Máquina* também possuem seus pressupostos filosóficos, e que diferentes tipos de *Inteligência de Máquina* são formados a partir de diferentes paradigmas. É difícil imaginar uma máquina de Turing linear apresentar algum tipo de comportamento ético que não seja baseado em normas pré-embutidas em seu código fonte. Porém quando analisamos um algoritmo de machine learning é muito mais fácil aceitar que o mesmo é capaz de aprender algum comportamento ético. Mas a moralidade é aprendida? Ou é um algo a priori pré-embutido no código fonte de nossos cérebros? Um pouco de cada? Nem os computadores acharam a resposta ainda.

James Moor (2006) em *The Nature, Importance, and Difficulty of Machine Ethics* discute a possibilidade de embutirmos comportamentos éticos nos agentes artificiais. Seu primeiro movimento é distinguir os diferentes agentes artificiais e categoriza-los de acordo com seu nível de autonomia e eficiência morais, tentando fundamentar se é possível a existência de um AMA.

Computadores convencionais são considerados meros *Agentes Normativos* (principalmente devido a sua origem linear/simbolista), e não necessariamente agentes éticos. Suas ações são éticas na medida que em que ações éticas corresponderem a ações eficientes. O segundo tipo é chamado de *Agentes de Impacto Ético*, que são aqueles que geram um efeito no mundo que seja idealizado como moralmente correto (porém sem capacidade de aprendizado).

Os *Agentes Implicitamente Éticos* são aqueles programados para causar um impacto positivo e evitar ações que causem impacto negativo (como o caso de um piloto robô para aviões, que não pode deixar o avião cair).

Se você embutir comportamentos éticos em uma máquina, como você faria? Alguém instruído nas matrizes analíticas da filosofia diria que basta programarmos as máximas éticas nessa máquina que seu comportamento será moralmente correto. Esta máquina conteria à priori tais máximas, desta forma elas fazem parte de sua natureza e se tornam parte indissociável da máquina, e o melhor, sem precisar de anos de treinamento e prática. Neste caso o

programador insere tais rotinas na máquina, e precisamos confiar na sua teória imparcialidade ao executar tal tarefa, pois a moralidade da máquina será a que sua equipe de desenvolvimento definir que é a mais adequada.

As categorias já citadas se enquadram de forma linear em agentes autônomos sem capacidade de aprendizado (*machine learning*), e podemos afirmar com certa propriedade que se encaixam dentro dos paradigmas Conexionista ou Simbolista. Já a última categoria (chamada de *Agentes Explicitamente Éticos*) engloba entes que são criados para calcular qual a melhor decisão de acordo com cada situação (clara influência Utilitarista). Isso obriga a máquina a ter capacidade de representar a realidade a sua volta, interpretá-la dentro de suas categorias pré-concebidas, e modelar suas ações de acordo a dinâmica do mundo externo.

Este é o primeiro caso onde temos algum nível (mesmo que baixo) de autonomia, pois a máquina se comporta baseada em informações que recebe. Longe de um livre arbítrio, mas agindo moralmente dentro de uma sociedade humana, e se adaptando a dinâmica da moral vigente. Existiria ainda uma quinta categoria (Agentes Completamente Éticos), que seriam máquinas com a mesma capacidade que a humana de tomar decisões. Logicamente que ainda não criamos uma máquina desse nível, devido ao fato de que para isso é necessário que a máquina tenha intencionalidade, consciência e liberdade, que são faculdades humanas.

Porém comportamento não significa consciência. Searle argumenta de forma decisiva: mesmo no *Machine Learning* as habilidades sintáticas ainda não foram capazes de fazer emergir capacidades semânticas. Mas se partimos do pressuposto que a moralidade das máquinas não precisa espelhar a moralidade humana, talvez o caminho para uma ética das máquinas não nos pareça tão dependente de inteligência. Não temos o código-fonte completo do ser humano para inserir nele uma moralidade “perfeita”, porém em um computador podemos ao menos tentar. A pergunta sem resposta seria sobre como essa moralidade deveria ser. Existem vários caminhos teóricos possíveis a discussão, e a nossa proposta foi tentar olhar de forma integral para os mais importantes, visando compreender como cada um deles, em seu pequeno universo delimitado, pode contribuir para o estudo filosófico da moralidade.

Assim sendo, as propostas de moralidade acerca das máquinas

precisam, no mínimo, levar em conta a qual tipo de paradigma pertence a máquina em questão, pois eles determinam os limites das ações de tais agentes. Tais limites se aplicam ao nível de pré-determinações no paradigma simbolista, e a capacidade de aprendizado no paradigma conexionista. Uma máquina simbolista linear necessita de uma moralidade pré-embutida (à priori) para que possa tomar alguma decisão que leve em conta o código moral vigente. Tal moralidade invariavelmente seria derivada se seus criadores ou idealizadores, de modo que dificilmente alguma das linhas téoricas da computação ou da filosofia iriam imputar responsabilidade num caso desses.

No que diz respeito ao paradigma conexionista, o machine learning permite as máquinas se adaptar ao meio, de modo que a moralidade vigente seria, mesmo que parcialmente, absorvida pela máquina no dia a dia. Um caso curioso diz respeito ao Tay, robô da Microsoft que se tornou racista e homofóbico analisando e repetindo as frases de pessoas reais que encontrava na internet através de data mining. Em menos de 24h a empresa o retirou do ar sob ameaça de processos e sérios problemas legais. A inteligência de máquina não pode sofrer algum tipo de processo, mas a Microsoft pode. Não se pode punir uma máquina, mas se pode exigir sob pena de punição a seus criadores que uma máquina desse porte não seja *ethical blind*.

Já uma crítica da moralidade das máquinas a partir de Heidegger, mesmo ele jamais tendo criado um explícito sistema ético, se torna possível a partir de uma perspectiva ontológica. As máquinas enfatizam o caráter apofântico do discurso, sem acessar a dimensão propriamente hermenêutica de significado da linguagem, fazendo com que até seu conhecimento seja isento de sentido.

Independente das linhas teóricas, me parece que se faz necessário uma ética pré-embutida aos agentes artificiais, que poderia ser complementada e afinada através de *machine learning*. Precisamos urgentemente fazer com que as ações de tais agentes obedeçam padrões morais que os orientem em situações de mesma dimensão, antes que sua emergente complexidade técnica torne tal tarefa ainda mais complexa.

CONCLUSÃO

Esta dissertação buscou responder a pergunta a cerca da delimitação do acesso e uso da moralidade por parte dos agentes artificiais. Dada a impossibilidade de imputação e a necessidade de evitarmos os problemas da criação de máquinas *ethically blind*, se faz necessário uma crítica que pergunte: até que ponto as inteligências de máquina podem acessar e fazer uso da moralidade?

Os agentes artificiais ainda não são imputáveis de responsablidade moral, porém seus criadores são potencialmente imputáveis em razão de danos causados por falhas nas suas criações, num caso de não haver preocupações de cunho ético no desenvolvimento de seu software. Tais preocupações se materializam na possibilidade de pré-embutir um código moral nas máquinas que são expostas a situações que exigem tomada de decisão de cunho ético. Porém as inteligências de máquinas são muito diferentes entre si, tanto ôntica quanto ontologicamente, de modo que não existe um caminho ideal e padronizado que possa ser instalado em todas elas para que seu agir se torne ético. Compreendendo abstratamente as bases de cada uma delas foi possível verificar quais melhor atendem aos pré-requisitos necessários para que se embutam instruções morais em seu código-fonte, ou quais possuem capacidade de aprender regras morais. Então até que ponto os agentes artificiais podem acessar e fazer uso da moralidade?

Para responder tal pergunta foi necessáriofundamentar filosoficamente o status ontológico de tais entes, buscando em Heidegger uma forma de compreender a formação dos paradigmas científicos que permitiram sua criação. Tendo identificado a visão científica hegemônica de base dos *agentes artificiais*, avançamos ao estágio de categorizar os tipos a partir de seu respectivo paradigma teórico, os relacionando com as linhas de pensamento filosófico correspondentes. Encontradas as demarcações ontológicas de cada paradigma, conseguimos compreender as limitações de tais agentes, concluindo que os mesmos não possuem acesso a dimensão semântica da linguagem, e que por isso seus critérios de imputabilidade moral não podem ser os mesmo utilizados com os seres humanos. Desta forma partimos em busca dos casos teóricos onde agentes artificiais se encontram perante

situações morais (AMAs), e encontramos o modelo de categorização de Moor, que nos permite diferenciar tais máquinas a partir de sua complexidade tecnológica, capacidade de aprendizado (*machine learning*) e capacidade de desempenho de suas designações.

Como as máquinas são criações do paradigma representacional, acabam por não escapar de suas inerentes limitações. Coube a este trabalho compreender suas bases ontológicas de modo que ficasse evidente a possibilidade ou não das mesmas acessar algum tipo de código moral. Para além da esfera do possível, concluímos que mais do que necessário, é urgente que pensemos e encontremos formas de fazer nossos agentes artificiais agir moralmente, pois chegamos num ponto onde delegamos aos mesmos situações complexas de nossa vida, que os colocam perante complicadas decisões morais que dizem respeito nossa vida.

Agora reunimos as condições necessárias para responder nosso problema sobre até que ponto os agentes artificiais podem acessar e fazer uso da moralidade?Talvez a resposta seja: depende do agente, da situação, e acima de tudo, do nosso entendimento sobre a moralidade. Usando a nomenclatura de Moor, *Agentes Normativos* (que são apenas eficientes) não possuem qualquer acesso a moralidade, devido a sua necessidade funcional. *Agentes de Impacto Ético*, *Implicitamente Éticos* e *Explicitamente Éticos* possuem um acesso, ainda que limitado a alguma camada de moralidade, na medida das necessidades para as quais foram projetados. Já os hipotéticos *Agentes Completamente Éticos*, devido ao fato de sua capacidade comparável a humana de tomada de decisões, teriam acesso a moralidade de uma maneria semelhante a nossa.

Tais agentes hipotéticos ainda não foram criados, mas nossa tecnologia está crescendo tão rapidamente que já não é mais tão certo afirmar que jamais serão criados. De nossa parte como humanidade, já deixamos a tecnologia de automação evoluir demasiado no que diz respeito a sua materialidade, mas seu software continua *ethically blind*, o que lentamente começa a nos causar alguns problemas. Mais de uma década atrás, Allen e Wallach chamaram atenção pro fato de que era necessário criar regras morais para nossas máquinas, e atéhoje avançamos muito pouco neste sentido, de modo que mesmo que tecnicamente já sejam capazes de atingir algum nível ético, nossas

inteligências de máquina, geralmente guiadas por interesses comerciais, continuam quase que totalmente vazias de moralidade.

REFERÊNCIAS

HEIDEGGER, Martin. Marcas do caminho. Tradução de Enio Paulo Giachini e Ernildo Stein. Petrópolis: Editora Vozes, 2008.

HEIDEGGER, Martin. O caminho para a linguagem. Trad. M. Scuback. In: A caminho da linguagem. Petrópolis, Vozes, 2003b. (Coleção Pensamento Humano).

HEIDEGGER, Martin. Os conceitos fundamentais da metafísica: mundo, finitude,solidão. Rio de Janeiro: Forense Universitária, 2003a.

HEIDEGGER, Martin. Que é Metafísica.In: HEIDEGGER. Os Pensadores.6 ed.Trad.: Ernildo Stein. São Paulo: Nova Cultural, 1996a. Schuback. Petrópolis: Vozes, 2005.

HEIDEGGER, Martin. Questions I et II. Paris: Galimard, 2006.

HEIDEGGER, Martin. Seminarios de Zollikon Tradução de Ángel Xolocotzi Yáñes. Ciudad de México: Herder, 2013.

HEIDEGGER, Martin. Ser e tempo I. Petrópolis: Vozes, 1989.

MCCARTHY, John and Others. “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence,” 1955. http://wwwformal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

MOOR, James. The Nature, Importance, and Difficulty of Machine Ethics. Hanover: Darthmout College, 2006.

NEWELL, A. eSIMON, H. Computer Science as Empirical Inquiry: Symbols and Search. Communications of the ACM, 19(3): 113–126, 1976.

SEARLE, John. “Minds, Brains, and Programs”. in The Behavioral and Brain Sciences 3. Berkeley: University of California, 1980.

SEARLE, John. The Rediscovery of the Mind, Cambridge: MIT, 1992.

SEARLE, John. Mente, cérebro e ciência. Lisboa: Edições 70, 1997.

SEARLE, J. R. (2006). A redescoberta da mente (E. P. e Ferreira, trans.). São Paulo: Martins Fontes. (Trabalho original publicado em 1992)

TURING, A.M. (1950). Computing machinery and intelligence. Mind, 59, 433-460.Oxford: Oxford Press, 1950.

Wallach, W., & Allen, C. (2009). Moral machines: Teaching robots right fromwrong. New York: Oxford University Press.