

**UNIVERSIDADE ESTADUAL PAULISTA – UNESP
FACULDADE DE FILOSOFIA E CIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA
CAMPUS DE MARÍLIA**

CAMILA DA CRUZ SILVA

**AGÊNCIA E RESPONSABILIDADE NO CONTEXTO DE SISTEMAS
AUTÔNOMOS ARTIFICIAIS**

**MARÍLIA
2020**

**UNIVERSIDADE ESTADUAL PAULISTA – UNESP
FACULDADE DE FILOSOFIA E CIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA
CAMPUS DE MARÍLIA**

CAMILA DA CRUZ SILVA

**AGÊNCIA E RESPONSABILIDADE NO CONTEXTO DE SISTEMAS
AUTÔNOMOS ARTIFICIAIS**

Dissertação apresentada à Faculdade de Filosofia e Ciências da Universidade Estadual Paulista - UNESP - Campus de Marília, como requisito parcial para obtenção do Título de Mestre em Filosofia. Área de concentração: Filosofia. Linha de pesquisa em Filosofia da Informação, da Cognição e da Consciência.

Orientadora: Prof.^a Dra. Mariana Claudia Broens

MARÍLIA
2020

S586a

Silva, Camila da Cruz

Agência e Responsabilidade no Contexto de Sistemas Autônomos
Artificiais / Camila da Cruz Silva. -- Marília, 2020

97 p.

Dissertação (mestrado) - Universidade Estadual Paulista
(Unesp), Faculdade de Filosofia e Ciências, Marília

Orientadora: Mariana Claudia Broens

1. Filosofia da tecnologia. 2. Tecnologia da informação. 3.
Sistemas artificiais. 4. Responsabilidade. 5. Ética da informação. I.
Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da
Faculdade de Filosofia e Ciências, Marília. Dados fornecidos pelo autor(a)

Essa ficha não pode ser modificada.

CAMILA DA CRUZ SILVA

**AGÊNCIA E RESPONSABILIDADE NO CONTEXTO DE SISTEMAS AUTÔNOMOS
ARTIFICIAIS**

Dissertação apresentada à Faculdade de Filosofia e Ciências da Universidade Estadual Paulista “Júlio de Mesquita Filho” – UNESP, Campus de Marília, como requisito parcial para obtenção do título de Mestre em Filosofia.

Área de Concentração: Filosofia da Mente, Epistemologia e Lógica.

Linha de pesquisa: Filosofia da Mente, Ciência Cognitiva e Semiótica.

BANCA EXAMINADORA:

Orientadora: Prof.^a Dr.^a Mariana Claudia Broens

Universidade Estadual Paulista “Júlio de Mesquita Filho”

Prof.^a Dr.^a Maria Eunice Quilici Gonzalez

Universidade Estadual Paulista “Júlio de Mesquita Filho”

Prof. Dr. Kleber Bez Birolo Candiotto

Pontifícia Universidade Católica do Paraná

Suplente interno: Prof. Dr. Kleber Cecon

Universidade Estadual Paulista “Júlio de Mesquita Filho”

Suplente externo: Prof. Dr. João Antonio de Moares

Faculdade João Paulo II (FAJOPA)

Marília – São Paulo, 18 de dezembro de 2020

AGRADECIMENTOS

Sou eternamente grata aos meus pais, Luciana e Olair, e à minha irmã, Carolina, que costumam ser o princípio de tudo em minha vida.

Agradeço à professora Mariana Broens, que com todo o seu cuidado e carinho nos dá coragem e forças para que sigamos, com amor e dedicação, os nossos trabalhos. Professora, gostaria de deixar registrada a minha admiração, você costuma ser a luz que ilumina o túnel todo.

Agradeço aos colegas João de Moraes e Ricardo Peraça, leitores da primeiríssima versão deste trabalho.

Agradeço à professora Maria Eunice e ao professor Kleber Candiottto pelo cuidado com o meu trabalho e comigo. Minha banca da Qualificação foi um dos momentos mais especiais da minha jornada acadêmica graças a vocês.

Agradeço aos colegas da faculdade e aos colegas do GAEC. Selmy, Marcus, Tauami, Sabrina, Bárbara Tavares, Bárbara Lulli, Valdirene, Bruna, Daphne, Felipe, Ana Paula 1 e 2, Douglas, entre outros tantos e ao mesmo tempo tão poucos, até mesmo o café mais despretensioso foi importante para que este momento se consolidasse.

Agradeço ao meu companheiro Rômulo por sempre estar ao meu lado lembrando-me de quem sou e encorajando-me a alcançar meus objetivos.

Aos amigos Ana Letícia, Nelson, Lizandra, Bruno, Bianca, vocês sempre aparecem em meu caminho quando menos espero, mas quando mais preciso.

Agradeço ao meu cunhado Alex e sua família: Zé, Arlene e Davi pelo carinho e acolhimento.

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil – (CAPES) – Código de financiamento 001.

“O que será que será?”

Chico Buarque (1976)

RESUMO

Esta pesquisa tem por objetivo investigar, no contexto das tecnologias informacionais emergentes, a possibilidade de sistemas autônomos artificiais serem considerados agentes e, em particular, agentes morais aos quais cabe atribuir algum grau de responsabilidade. A pergunta que guiará esta investigação pode ser assim formulada: Seria legítimo atribuir algum grau de responsabilidade a sistemas autônomos artificiais, levando em consideração que estes sistemas são produzidos e programados por sistemas autônomos naturais e intencionais? Para tanto, explicitaremos se, e em que condições, podemos classificar um agente artificial como autônomo e o que entendemos como conduta responsável. Procuraremos mostrar que sistemas artificiais que forem capazes de extrapolar sua programação inicial, aprender com a própria experiência e alterar o rumo de sua ação como resultado de tal aprendizagem podem vir a ser considerados elementos constituintes de uma rede de agentes intencionais. Tal atribuição estaria baseada na tese de que agentes artificiais, instanciadores de intencionalidade derivada, teriam também corresponsabilidade derivada sobre as possíveis consequências de suas *performances*. Por outro lado, recusar a possibilidade de agentes artificiais poderem ser considerados corresponsáveis pelas consequências de suas performances implicaria não considerar devidamente a capacidade de tais sistemas alterar o rumo de sua conduta como resultado de sua própria experiência. Assim, entendemos que a clássica relação entre agência e responsabilidade moral sobre as consequências da conduta precisa ser revista de modo a incluir a hipótese de agência artificial e suas possíveis relações com a responsabilização moral no contexto das tecnologias contemporâneas.

Palavras-chave: Filosofia da tecnologia. Tecnologias informacionais. Sistemas artificiais. Responsabilidade. Ética da informação.

ABSTRACT

This research aims to investigate, in the context of emerging information technologies, the possibility of artificial autonomous systems considered to be agents and, in particular, moral agents to which it is appropriate to attribute some degree of responsibility. The question that will guide this investigation can be formulated as follows: Would it be legitimate to attribute some degree of responsibility to artificial autonomous systems, taking into account that these systems are sought and programmed by natural and intentional autonomous systems? For that, we will explain if, and under what conditions, we can classify an artificial agent as autonomous and what we understand as responsible conduct. We will try to show that artificial systems that enable them to extrapolate their initial programming, learn from their own experience and change the course of their action as a result of such learning can come to be considered as constituent elements of a network of intentional agents. This must be based on the possibility that the artificial agents, instantiators of derived intentionality, also derived co-responsibility for the possible consequences of their performances. On the other hand, to refuse the possibility that artificial agents can be considered co-responsible for the consequences of their performances, would imply that we do not properly consider the capacity of such systems to change the course of their conduct as a result of their own experience. Thus, we understand that the classic relationship between agency and moral responsibility on the consequences of conduct needs to be revised to include the hypothesis of artificial agency and its possible relationships with moral accountability in the context of contemporary technologies.

Keywords: Philosophy of technology. Information technologies. Artificial systems. Responsibility. Information ethics.

SUMÁRIO

INTRODUÇÃO.....	8
CAPÍTULO 1 - ANÁLISE DO CONCEITO DE AGÊNCIA.....	11
Apresentação.....	12
1.1 Análise do conceito de agência	12
1.2 Diferentes tipos de agência.....	20
1.3 Agência e autonomia	24
1.4 Agência e autonomia em sistemas artificiais.....	29
CAPÍTULO 2 - CONDUTA MORAL E TIPOS DE RESPONSABILIDADE.....	35
Apresentação	36
2.1 Atribuição de responsabilidade	37
2.2 Responsabilidade coletiva e ações em conjunto.....	41
2.3 Conduta moral responsável e novas tecnologias de informação	45
2.4 Sistemas artificiais podem ser considerados agentes morais?.....	53
CAPÍTULO 3 - MÁQUINA CRIANÇA E GRAUS DE AUTONOMIA.....	60
Apresentação.....	61
3.1 - Máquina criança	61
3.2 - Máquinas cometem erros?.....	72
3.3 - Graus de autonomia.....	79
CONSIDERAÇÕES FINAIS	88
REFERÊNCIAS	93

INTRODUÇÃO

O objetivo geral deste trabalho consiste em investigar a possibilidade de certos sistemas artificiais, ditos autônomos, poderem ser considerados agentes em um sentido não trivial e, em particular, agentes morais aos quais cabe atribuir algum grau significativo de responsabilidade. A pergunta que guiará esta investigação pode ser assim formulada: Seria legítimo atribuir algum grau de responsabilidade a sistemas autônomos artificiais, levando em consideração que estes sistemas são produzidos e programados por sistemas autônomos naturais e intencionais?

Ao tratarmos de tecnologia nos dias atuais, parece natural pensarmos em sistemas altamente elaborados e sofisticados. Porém, a técnica e a tecnologia estão presentes na vida e no cotidiano das pessoas desde a primeira ferramenta construída. O que vivenciamos na atualidade pode ser considerado apenas uma continuidade dos movimentos iniciados por nossos ancestrais, embora os resultados desse processo de elaboração de tecnologias tenham se complexificado e adquirido um elevado grau de sofisticação. Desse modo, a tecnologia envolveria tanto o registro na fabricação de ferramentas, quanto o aprimoramento dessas ferramentas, como indica Kurzweil (2007, p. 34).

Nesse sentido de progressão e sofisticação, observamos que os seres humanos, assim como muitos outros animais, buscam maneiras de adaptar seu ambiente às suas necessidades, e se atualmente nos encontramos imersos em um ambiente no qual dispositivos eletrônicos cumprem diversas tarefas, antes reservadas apenas às pessoas, é porque o atual desenvolvimento da sociedade proporciona essa possibilidade.

Tendo como suposto objetivo a praticidade, o conforto, a economia de tempo e outros fatores que permitam que as pessoas dediquem esforços para atividades que julguem mais importantes e significativas, começamos a construir e a usar artefatos tecnológicos cujas funcionalidades frequentemente fogem do controle e da compreensão da maioria de seus usuários. Assim, nos vemos cercados por tecnologias altamente sofisticadas, as quais, além de executar diversas tarefas de ordem prática, desempenham também um papel mediador em nossas relações interpessoais.

Desse modo, após um longo período de desenvolvimento e aperfeiçoamento de ferramentas, o nosso contexto atual é caracterizado pela computação ubíqua (WEISER, 1991) que consiste na computação em rede integrada 24 horas no ambiente e que proporciona

inúmeras diferentes funcionalidades a partir do uso de artefatos eletrônicos sofisticados que possibilitam a realização de atividades que antes exigiam presencialidade.

Com o surgimento das tecnologias ubíquas de informação e comunicação (TIC) portáteis ligadas na rede Internet, as pessoas, a partir de seus *smartphones*, *smartwatches*, *tablets* e computadores pessoais, para citar alguns dos exemplos mais comuns, passaram a resolver de forma mais fácil e rápida muitas de suas questões cotidianas.

Assim, de acordo com as considerações realizadas até o momento, entendemos que a era informacional digital é marcada pela expansão globalizada dos processos informacionais e comunicativos no ciberespaço, no qual a capacidade de coleta, armazenamento, mineração e análise de dados é cada vez mais ampla.

O meio digital permite que pessoas de quase qualquer parte do mundo se mantenham conectadas 24 horas por dia através da Internet e de seus dispositivos eletrônicos. Essas tecnologias fascinam e envolvem os usuários uma vez que o meio digital proporciona a possibilidade de receber e compartilhar informações advindas de qualquer parte do globo, a resolução rápida de problemas de ordem prática, além de proporcionar a conexão a qualquer outro usuário que faça uso das TIC, para assim compartilhar momentos, ideias e culturas.

De forma resumida, a era informacional digital pode ser caracterizada enquanto mudança das atividades comuns das pessoas para o espaço digital, um espaço vasto e cheio de possibilidades e com poucas restrições, cujas leis ainda não conhecemos, e ao qual nossas leis e normas sociais não podem ser facilmente aplicadas por, talvez ainda, não ser possível estabelecer limites claros neste ambiente, especialmente no que tange a atribuição de responsabilidades por eventuais ações danosas.

Levando em consideração o panorama acima traçado, esta pesquisa está situada no contexto das TIC, da sociabilidade marcada pelos ambientes digitais e do desenvolvimento de artefatos tecnológicos bem como das consequências de suas aplicações para os seres humanos e o ambiente em geral. Entendemos que o desenvolvimento e a evolução de artefatos tecnológicos e da computação ubíqua, assim como o aprimoramento da programação que permite que essas tecnologias atuem em conjunto com os seres humanos, tem como objetivo atingir um grau tão elevado de automação que possibilite que determinadas tecnologias venham a se tornar autônomas na medida em que agem e deliberam sem intervenção direta de seus idealizadores, passando a ser em algum grau independentes de seus criadores.

Nosso trabalho consiste, assim, em investigar a possibilidade de sistemas artificiais gestados nesse ambiente tecnológico-informacional serem portadores de algum grau de agência e autonomia. Consequentemente, pretendemos investigar se a possibilidade de

atribuição de responsabilidade a sistemas artificiais se justifica, pois, no que diz respeito a artefatos tecnológicos altamente sofisticados, portadores de algum grau de agência e autonomia, não é possível prever seus *outputs* e suas consequências para a vida humana.

Assim, tendo tal objetivo geral, no **Primeiro Capítulo: Análise do conceito de agência**, investigaremos os conceitos de ‘agência’ e de ‘autonomia’ para verificar se é possível considerar que certas tecnologias sejam portadoras de agência e de autonomia e em quais sentidos e contextos a atribuição dessas características pode ser justificada.

No **Segundo Capítulo: Conduta moral e responsabilidade coletiva**, investigaremos os conceitos de ‘responsabilidade’ e de ‘conduta moral responsável’, abordando diferentes tipos de responsabilidade e situações nas quais seria possível justificar a atribuição de responsabilidade a sistemas artificiais.

No **Terceiro Capítulo: Máquina criança e graus de autonomia**, será apresentada uma caracterização de autonomia não mais em sentido absoluto (de ter ou não ter autonomia), mas de autonomia-processo, isto é, da possibilidade de aquisição gradativa de autonomia em uma escala, a partir da qual seria possível, igualmente, atribuir graus de responsabilidade a um agente. Para tanto, exploraremos o experimento de pensamento sugerido por Turing (1950) sobre a possibilidade de construção de uma “máquina criança”, investigaremos a questão de máquinas inteligentes cometerem erros (especialmente crucial quando se trata da possibilidade, já em implementação experimental, de veículos autônomos provocarem acidentes) e por fim, iniciaremos a construção de uma linha de raciocínio que nos permitirá identificar diversos graus de autonomia de sistemas artificiais.

Nos comentários finais teceremos considerações a respeito das análises propostas ao longo desta pesquisa a fim de analisarmos algumas das implicações dos resultados que obtivemos ao longo do texto.

CAPÍTULO 1

ANÁLISE DO CONCEITO DE AGÊNCIA

APRESENTAÇÃO

No presente capítulo investigaremos os conceitos de ‘agência’ e de ‘autonomia’ para verificar se é possível considerar que certas tecnologias sejam portadoras de agência e de autonomia e em quais sentidos e contextos a atribuição dessas características pode ser justificada. Para tanto, na seção **1.1 Análise do conceito de agência**, faremos uma caracterização do conceito de “agência” a partir da Filosofia da Mente e da Ação contemporâneas. Na seção **1.2 Diferentes tipos de agência** discutiremos a caracterização de tipos de agência que não se enquadram nas descrições feitas na primeira seção. Na seção **1.3 Agência e autonomia** investigaremos em quais condições é possível atribuir autonomia a um agente. Por fim, na seção **1.4 Agência e autonomia em sistemas artificiais** investigaremos em quais contextos sistemas artificiais poderiam, ou não, ser considerados portadores de agência e em quais contextos esses agentes artificiais poderiam ser considerados agentes autônomos em um sentido não metafórico.

1.1 Análise do conceito de agência

Para dar início ao tópico que tratará sobre a agência autônoma em sistemas naturais e artificiais, é fundamental introduzir partes essenciais do debate filosófico realizado para esclarecer o conceito de agência. O estudo desse conceito gerou diversas interpretações, considerando que, em sentido geral, conforme Markus Schlosser (2019) indica, é possível encontrar agência em todo lugar. “Sempre que as entidades entram em relações causais, pode-se dizer que elas agem umas sobre as outras e interagem entre si, provocando mudanças umas nas outras.” (SCHLOSSER, 2019, p. 2. Tradução nossa).¹

Desse modo, dada a dificuldade em definir o conceito de agência e a admissão de agência em diversas e distintas entidades, procuraremos verificar a possibilidade de sistemas artificiais serem considerados como instanciadores de agência em um sentido forte.

Enfatizamos ainda que os conceitos de *agência*, *agente* e *ação* são interligados e mutuamente dependentes na medida em que a *agência* é propriedade do *agente* que pratica ou realiza *ações* no mundo.

Segundo Schlosser (2019), os estudos sobre agência foram relacionados aos estudos sobre a ação intencional. Essa forma de relacionar os dois temas pode ser encontrada em

¹ Whenever entities enter into causal relationships, they can be said to act on each other and interact with each other, bringing about changes in each other.

trabalhos de figuras históricas da Filosofia, tais como Aristóteles e Hume, e na filosofia analítica contemporânea, sendo os trabalhos mais influentes sobre os conceitos de ação intencional e de agência remetidos à Elizabeth Anscombe (1975) e a Donald Davidson (1963).

A Filosofia da Ação contemporânea fornece uma “concepção padrão” da ação e uma “teoria padrão” da ação. Anscombe (1975) - responsável pela “concepção padrão da ação” - analisa e interpreta a ação intencional. E Davidson (1963) - responsável pela “teoria padrão da ação” - explica a intencionalidade da ação nos termos de uma causação decorrente de estados mentais e de eventos relacionados ao agente.

Porém, como indica Schlosser (2019, p. 3), os debates acerca do conceito de agência criaram certa resistência e deixaram de lado sua relação com a ação intencional, promovendo assim concepções alternativas a respeito do conceito.

Antes de explorarmos concepções alternativas de agência, julgamos ser importante analisar a concepção de agência considerada padrão na Filosofia da Mente e da Ação contemporâneas, apresentando suas principais características. Schlosser esclarece que:

No centro da concepção padrão estão as duas reivindicações a seguir. Primeiro, a de que a noção de ação intencional é mais fundamental que a noção de ação. Em particular, a ação deve ser explicada nos termos da intencionalidade da ação intencional. Em segundo lugar, há uma conexão íntima entre ação intencional e agir por uma razão. (SCHLOSSER, 2019, p. 3. Tradução nossa)²

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

Como veremos adiante, para Davidson (1963, 1971), uma ação será sempre intencional, isso significa que a ação poderá sempre ser explicada de acordo com uma razão do ponto de vista do agente. Por sua vez, um acontecimento que não possa ser explicado ou racionalizado pelo agente será considerado como sendo um evento.

Segundo Davidson (1963, p. 685), a racionalização de uma ação é uma espécie de explicação causal ordinária. Para o autor, essa racionalização só acontece se o agente permitir que outras pessoas vejam qualquer característica que o mesmo tenha visto, ou que pensou ter visto, para executar a sua ação. Essas características incluem quaisquer consequências ou

² At the core of the standard conception are the following two claims. First, the notion of intentional action is more fundamental than the notion of action. In particular, action is to be explained in terms of the intentionality of intentional action. Second, there is a close connection between intentional action and acting for a reason.

aspectos que indiquem o motivo pelo qual o agente tenha executado determinada ação. Esse motivo pode ser reconhecido como um desejo do agente, ou como o processo meditativo e reflexivo geralmente realizado antes da tomada de decisões, o motivo pode ser ainda algo que beneficia ou obriga o agente a agir de determinada maneira.

Uma razão racionaliza uma ação apenas se nos leva a ver algo que o agente viu, ou pensou ter visto, em sua ação - alguma característica, consequência ou aspecto da ação que o agente desejou, valorizou, considerou, zelou, achou vantajoso, obrigatório ou aceitável. Não podemos explicar por que alguém fez o que fez simplesmente dizendo que a ação em particular o atraiu; devemos indicar o que foi sobre a ação que recorreu. (DAVIDSON, 1963, p. 685. Tradução nossa)³

O que significa então dizer que alguém agiu por determinado motivo, ou que uma ação tem uma razão que a sustenta? De acordo com Davidson, um agente que racionaliza sua ação pode ser classificado como “[...] (a) tendo alguma *pro attitude* que direciona a ações de um certo tipo, e (b) acreditando (ou sabendo, percebendo, notando, lembrando) que sua ação é daquele tipo.” (DAVIDSON, 1963, p. 685. Tradução nossa).⁴

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

Aqui a palavra 'atitude' cumpre serviço de subordinada, pois deve abranger não apenas traços permanentes de caráter que se mostram em uma vida inteira de comportamento, como amor por crianças ou gosto por companhia elevada, mas também a fantasia mais passageira que leva a uma ação única, como um desejo repentino de tocar o cotovelo de uma mulher. Em geral, *pro attitudes* não devem ser tomadas por convicções, ainda que temporárias, de que toda ação de um certo tipo a ser realizada, vale a pena ser realizada (...). (DAVIDSON, 1963, p. 686. Tradução nossa)⁵

³ A reason rationalizes an action only if it leads us to see some- thing the agent saw, or thought he saw, in his action-some feature, consequence, or aspect of the action the agent wanted, desired, prized, held dear, thought dutiful, beneficial, obligatory, or agree- able. We cannot explain why someone did what he did simply by saying the particular action appealed to him; we must indicate what it was about the action that appealed.

⁴ [...] (a) having some sort of *pro attitude* toward actions of a certain kind, and (b) believing (or knowing, perceiving, noticing, remembering) that his action is of that kind.

⁵ The word ‘attitude’ does yeoman service here, for it must cover not only permanent character traits that show themselves in a lifetime of behavior, like love of children or a taste for loud company, but also the most passing fancy that prompts a unique action, like a sudden desire to touch a woman’s elbow. In general, *pro attitudes* must not be taken for convictions, however temporary, that every action of a certain kind ought to be performed, is worth performing (...).

A essa relação causal, ou seja, a relação que implica encontrar um motivo que racionalize e explique por que o agente agiu de determinada maneira, Davidson dá o nome de razão primária (*primary reason*). Nas palavras do autor:

1. Para entendermos como uma razão de qualquer tipo racionaliza uma ação é necessário e suficiente que vejamos, ao menos em um esboço essencial, como construir uma razão primária.
2. A razão primária para uma ação é sua causa. (Davidson, 1963, p. 686. Tradução nossa)⁶

Assim, as razões primárias, que justificam a execução de ações, consistem em uma *pro attitude* e em uma crença relacionada da maneira correta. Para entendermos o significado do que vem a ser uma crença e uma *pro attitude* corretamente relacionadas trazemos as palavras de Jeff Speaks (2009), que utiliza o clássico exemplo de Davidson, no qual o agente acende a luz e acaba por alertar um ladrão que se encontra nos arredores.

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

O exemplo do ladrão; por que dizemos que o agente ligar o interruptor foi intencional, mas que ele alertar o ladrão não foi? Parece quase impossível não dizer que a diferença é que o agente queria (ou seja, tinha uma certa atitude proativa) acender a luz, mas não queria alertar um ladrão; ou, novamente, que o agente sabia que sua ação acenderia a luz, mas não sabia que sua ação alertaria o ladrão. Reflexões sobre casos como esse parecem deixar claro que a diferença entre atos intencionais e não intencionais pode ser encontrada nas crenças e desejos do agente. (SPEAKS, 2009, p. 2. Tradução nossa)⁷

É neste sentido que as razões primárias são justificativas para uma ação. Com o exemplo mencionado acima entendemos que a *pro attitude* ilustrada no ato de acionar o interruptor, e a crença de que o fazendo a luz se acenderá e iluminará o ambiente, a ação está

⁶ 1. For us to understand how a reason of any kind rationalizes an action it is necessary and sufficient that we see, at least in essential outline, how to construct a primary reason.

2. The primary reason for an action is its cause.

⁷ The example of the burglar; why do we say that the agent's turning on the light switch was intentional, but that his alerting the burglar was not? It seems almost impossible not to say that the difference is that the agent wanted (i.e., had a certain pro-attitude) to turn on the light, but did not want to alert a burglar; or, again, that the agent knew that his action would turn on the light, but did not know that his action would alert the burglar. Reflection on cases like this seems to make it clear that the difference between intentional and unintentional acts is to be found in the beliefs and desires of the agent.

racionalmente justificada, é uma ação intencional. Porém, como acionar o interruptor não se relaciona com o alertar o invasor, essa ação não é intencional no sentido de que não existe uma crença que se alinhe ao desejo do agente.

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

É geralmente aceito que uma explicação da razão de uma ação geralmente torna a ação inteligível ao revelar a meta ou intenção do agente. De acordo com teorias não causais, ter objetivos ou intenções relevantes não consiste na posse de estados ou eventos mentais causalmente eficazes. (SCHLOSSER, 2019, p. 6. Tradução nossa)⁸

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

Ainda que a ideia central da concepção padrão da ação expresse que as ações devem ser explicadas nos termos da intencionalidade e da ação intencional, tanto Davidson quanto Anscombe contribuíram para a formulação da concepção padrão da ação. Segundo Schlosser (2019, p.4), nessa formulação da ação a “(...) ação intencional é mais fundamental que a ação por si mesma: ações derivam e são dependentes da ação intencional”. (Tradução nossa)⁹

O que diferenciaria, então, a concepção padrão da ação da teoria padrão da ação? Davidson foi além da concepção padrão da ação ao propor que a teoria padrão da ação fornece explicações causais da ação e explicações baseadas na razão, de forma que, para Davidson, a razão opera causalmente (1963). É por este motivo que muitas vezes essa teoria é chamada de “teoria causal da ação”, como enfatiza Schlosser:

Esta teoria diz, aproximadamente, que algo é uma ação intencional e que foi feito por razões apenas no caso de ter sido causado por estados mentais e eventos certos de maneira certa. Os estados mentais certos e os eventos são estados e eventos que racionalizam a ação a partir do ponto de vista do

⁸ It is generally agreed that a reason explanation of an action usually renders the action intelligible by revealing the agent's goal or intention. According to non causal theories, having the relevant goals or intentions does not consist in the possession of causally efficacious mental states or events.

⁹ (...) intentional action is more fundamental than action itself: action derives from and is dependent on intentional action.

agente (tais como desejos, crenças e intenções). (...). Falando estritamente, esta é uma teoria evento-causa e consiste em uma teoria evento-causa da explicação racionalizada e em uma teoria evento-causa da ação intencional. Em conjunto com a concepção padrão, esta teoria causal nos fornece uma teoria da ação, que tem sido a teoria padrão na filosofia contemporânea da mente e da ação. (2019, p. 5-6. Tradução nossa)¹⁰

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

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

Mas, de acordo com Anscombe (1957), se tentarmos descrever e explicar o conceito de 'intenção' considerando apenas um dos casos citados como sendo suficiente para explicá-lo em todos os demais casos, estaremos utilizando-o equivocadamente:

[...] podemos dizer que a 'Intenção sempre diz respeito ao futuro'. Mas uma ação pode ser intencional sem se preocupar de forma alguma com o futuro. Perceber isso pode nos levar a dizer que há vários sentidos da [palavra] 'intenção', e que talvez seja completamente enganoso dizer que a palavra 'intencional' necessariamente está conectada com a palavra 'intenção', uma vez que uma ação pode ser intencional sem que contenha alguma intenção. Ou, como alternativa, podemos ficar tentados a pensar que apenas ações realizadas com certas intenções adicionais devem ser chamadas de intencionais. (ANSCOMBE, 1957, p. 1. Tradução nossa)¹¹

Para Anscombe, embora seja tentador afirmar que o conceito de intenção contenha em si diversos sentidos, é inadmissível dizer que tal conceito seja ambíguo, isto é, que possua diferentes acepções enquanto se apresenta em diferentes situações. (1957, p. 1).

¹⁰ This theory says, very roughly, that something is an intentional action and done for reasons just in case it is caused by the right mental states and events in the right way. The right mental states and events are states and events that rationalize the action from the agent's point of view (such as desires, beliefs, and intentions). (...) Strictly speaking, it is an event-causal theory and it consists of an event-causal theory of reason explanation and an event-causal theory of intentional action. In conjunction with the standard conception, this causal theory provides us with a theory of action, which has been the standard theory in the contemporary philosophy of mind and action.

¹¹ [...] we might say 'Intention always concerns the future'. But an action can be intentional without being concerned with the future in any way. Realizing this might lead us to say that there are various senses of 'intention', and perhaps that it is thoroughly misleading that the word 'intentional' should be connected with the word 'intention', for an action can be intentional without having any intention in it. Or alternatively we may be tempted to think that only actions done with certain further intentions ought to be called intentional.

Como destaca Rachel Wiseman, em *Anscombe's Intention* (1957/2016), a investigação de Anscombe sobre o conceito de 'intenção' se inicia a partir da pergunta: "O que distingue ações intencionais daquelas que não são?" (p. 78. Tradução nossa)¹²

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

A resposta de Anscombe é o que Wiseman (2016) chama de "voz descritiva", e nada daquilo que se segue à resposta deverá ser posto sob suspeição, de forma que as descrições das razões para uma ação não devem ser contestadas. Porém, para Anscombe, a resposta sugerida para iniciar sua investigação não é satisfatória uma vez que ao questionarmos o sentido relevante do 'Por que?' e o que significa dar uma razão para o agir, falamos sobre a mesma coisa.

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

Ações intencionais são praticadas por alguma razão de forma que a resposta do agente à pergunta 'por que realizou X' deve dar um sentido à sua ação, ou seja, o significado de uma ação intencional deve ser capaz de ser explicado através da resposta ao porquê o agente realizou X ou Y. Assim a intenção não pode ser confundida com a causa de uma ação, uma vez que, de acordo com Duncan Richter, a causa de uma ação não explica seu efeito e as ações intencionais não podem ser separadas ou diferenciadas das intenções presentes nestas ações:

[...] imagine um alpinista que perde a vontade de viver e, assim, solta a rocha e cai. Essa foi uma ação intencional, um ato de suicídio. Agora imagine que ele simplesmente perde o controle e cai. Essa não é uma ação intencional e não é suicídio, mas um acidente trágico. Agora imagine que ele tenha um impulso suicida momentâneo. Chocado com esse pensamento, ele perde a concentração e se solta, caindo da rocha. Essa queda foi causada por um pensamento de suicídio, talvez até por uma intenção de cometer suicídio, mas ainda assim foi um acidente, causado por um choque, e não por deliberação. (RICHTER, 2020. Tradução nossa)¹³

¹² What distinguishes actions which are intentional from those which are not?

¹³ [...] imagine a climber who loses the will to live and so let go of the rock and falls to his death. This was intentional, an act of suicide. Now imagine that he simply loses his grip and falls. This is unintentional and not suicide but a tragic accident. Now imagine that he has a momentary suicidal impulse. Shocked by this thought he loses concentration and let go, falling to his death. This fall was caused by a thought of suicide, perhaps even by an intention to commit suicide, but it was an accident still, caused by a shock rather than carried out deliberately

Neste sentido, de acordo com Richter, intenções não são a causa de ações intencionais, intenções não são eventos anteriores ao ato, intenções são uma parte ou aspecto da ação, ou seja, intenções são aquilo que revela à pergunta ‘por que o agente fez isso ou aquilo?’: o propósito do agente e o que significa sua ação.

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

Anscombe escreveu seu trabalho *Intention* para tentar explicar as motivações que levam um agente a praticar determinada ação. Segundo a filósofa, seu propósito original era mostrar que a concessão do título de Doutor *Honoris causa* da Universidade de Oxford ao ex-presidente dos Estados Unidos, Harry Truman, basicamente por ter lançado as bombas atômicas nas cidades japonesas de Hiroshima e Nagasaki em 1945, seria moralmente inaceitável. Com esta decisão, Truman consentiu com a morte de mais de duzentos mil civis, a fim de defender interesses dos aliados e em nome de um bem maior: o fim da guerra e a vitória. Segundo John Haldane (2020):

Perplexa pelos defensores de Truman, [Anscombe] chegou à conclusão de que eles não compreenderam a natureza de suas ações, e foi isso que a levou a escrever *Intention*, obra na qual ela apontou que, ao fazer uma coisa (mover a mão), pode-se intencionalmente estar fazendo outra (levando à morte de seres humanos). (Tradução nossa)¹⁴

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

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

¹⁴ Perplexed by defenders of Truman she came to the conclusion that they failed to understand the nature of his actions, and it was this that led her to write *Intention*, in which she pointed out that in doing one thing (moving one's hand) one may intentionally be doing another (directing the death of human beings).

Porém, como afirma Schlosser (2019), parece que existem seres que são capazes de agência genuína e que não possuem estados mentais representacionais como os dos seres humanos e quando o foco não é mais a agência humana, outros desafios emergem. Schlosser 2019, p.16 indica ainda que existem outros candidatos capazes de diferentes tipos de agência que não requerem representações mentais:

Eles incluem agência mental, agência compartilhada, agência coletiva, agência relacional e agência artificial. Em cada caso, podemos perguntar se a agência em questão pode ser explicada em termos da concepção e teoria padrão, ou se é de fato um tipo diferente de agência. (Tradução nossa)¹⁵

No presente tópico, nosso objetivo foi analisar o conceito de agência em debates filosóficos tanto tradicionais quanto contemporâneos. Vimos que as principais teorias aqui tratadas apresentam a intencionalidade como requisito para o reconhecimento da agência e, nesse sentido, apenas seres humanos com capacidade para justificar, reconhecer, racionalizar suas ações, seriam agentes. Essas capacidades mentais envolvem, em parte, aquilo que na filosofia da mente é compreendido como representações mentais, e para as representações mentais acontecerem seria necessário que o agente tenha capacidade de formação de conteúdo proposicional.

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

1.2 Diferentes tipos de agência

Como colocado anteriormente, parecem existir diversos tipos de agência que podem ser reconhecidos enquanto agência genuína, mas que não se encaixam nas determinações da teoria e da concepção padrão da agência. Além da assim considerada teoria padrão da agência caracterizada na seção anterior Anscombe (1957) e Davidson (1963), Schlosser menciona em seu trabalho mais dez tipos de agência. Os três tipos de agência que Schlosser julga como sendo tipos refinados são: agência autocontrolada, agência autônoma e agência livre. Eles se diferenciam dos demais, pois os outros tipos de agência são considerados mais básicos, ou seja, não requerem atribuição de estados mentais representacionais (SCHLOSSER, 2019).

Considerando a classificação proposta por Schlosser (2019), entendemos que, ao tratar de agência livre, agência autocontrolada e agência autônoma, o filósofo está tratando de

¹⁵ They include mental agency, shared agency, collective agency, relational agency, and artificial agency. In each case, we can ask whether the agency in question can be explained in terms of the standard conception and theory, or whether it is indeed a different kind of agency.

tipos de agência que envolvem representações mentais para sua efetivação. Mais adiante em nossa investigação trataremos tanto da autonomia de agentes naturais, quanto da autonomia de sistemas artificiais, e veremos que, ainda que um sistema artificial não possua a vida mental de um agente natural, por exemplo, de um agente humano, este também pode ser considerado como um agente autônomo em algum grau.

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

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

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

Assim, o ato mental de se pensar que *p* seria uma ação intencional apenas se o agente tiver uma intenção que inclua este pensamento em *p*. Esta tese é problemática uma vez que seria necessário ter outra intenção de se pensar em um determinado pensamento antes de

¹⁶ First, it seems that such mental occurrences are hardly ever, if ever, intentional actions. According to the standard theory, an event is an intentional action of the type A only if the agent has an intention that includes A in its content. In the basic case, this would be an intention to A. In an instrumental case, this would be an intention to perform some other action B in order to A. [...] Take the thought that *p*. According to the standard theory, thinking that *p* is an intentional action only if the agent has an intention that includes "think that *p*" in its content. This is rather odd and problematic, because we would have to have the intention to think a certain thought before we think it. Second, there are problems with the central case of decision-making.

pensá-lo, podendo-se cair, inclusive, numa regressão infinita. Por sua vez, o segundo problema apontado por Schlosser (2019, p. 17) se refere a casos de tomada de decisão:

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

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

A agência epistêmica, por sua vez, diz respeito ao controle que os agentes podem exercer sobre suas crenças. Distingue-se, nesta consideração de agência, duas posições principais: voluntarismo opinativo indireto e voluntarismo opinativo direto.

O voluntarismo indireto se refere às formas a partir das quais se é possível adquirir ou revisar crenças fazendo pesquisa, conduzindo experimentos, buscando legitimar hipóteses, analisando argumentos, avaliando evidências e considerando opiniões opostas. De acordo com Schlosser (2019, p. 19): “é indubitável que possamos exercer controle sobre nossas crenças de maneira indireta.” (Tradução nossa)¹⁸.

Já o voluntarismo opinativo direto supõe ser possível ter controle voluntário direto sobre algumas crenças. Segundo Schlosser (2019, p. 18), esse controle voluntário é normalmente compreendido como o tipo de controle que os agentes exercem quando praticam ações intencionais:

A questão principal aqui é que o voluntarismo opinativo direto parece ser incompatível com a natureza das crenças. Supõe-se que as crenças devem representar o mundo (ou “ter como meta a verdade”). Pode-se argumentar que não há diferença fundamental no controle sobre a ação e sobre a formação de crenças, porque em ambos os casos o controle consiste

¹⁷ According to the standard theory, deciding to A would be an intentional action only if one already had the intention to make a decision that includes “deciding to A” in its content. This seems, again, rather odd and problematic. Further, our reasons for making a decision to A are usually our reasons to A—they are reasons for performing the action. According to the standard theory, something is an action only if it has a reason explanation (in terms of the agent’s desires, beliefs, and intentions). As reasons are usually reasons for action, it is again difficult to see how making a decision can ever be an action.

¹⁸ It is fairly uncontroversial that we can exercise control over our beliefs in such indirect ways.

basicamente na capacidade de resposta da razão. Mas essa proposta ignora o papel central das intenções. (Tradução nossa)¹⁹

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

Schlosser (2019, p. 19) caracteriza a agência compartilhada de forma breve afirmando que esta ocorre quando dois ou mais indivíduos praticam uma ação em conjunto, como exemplo podemos pensar em casos em que duas ou mais pessoas cantam uma canção ou carregam juntos algum objeto.

A agência coletiva, por sua vez, acontece quando dois ou mais indivíduos agem como uma organização, “(...) de acordo com certos princípios ou procedimentos que constituem e organizam o grupo” (SCHLOSSER, 2019, p. 19. Tradução nossa)²⁰.

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

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

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

¹⁹ A main issue here is that direct doxastic voluntarism appears to be incompatible with the nature of beliefs. Beliefs are supposed to represent the world (or “aim at truth”). One may argue that there is no fundamental difference in the control over action and belief-formation, because in both cases the control consists basically in reason-responsiveness. But this proposal overlooks the central role of intentions.

²⁰ (...) in accordance with certain principles or procedures that constitute and organize the group.

autônomo, de um modelo individualista para outro que incorpora o contexto social. (Tradução nossa)²¹

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

Apresentamos, por fim, o último tipo de agência da lista de Schlosser (2019), a agência artificial. De acordo com algumas considerações sobre a possibilidade de agência em sistemas artificiais, este tipo de agência não pode ser justificado, uma vez que sistemas artificiais não apresentam estados internos que seriam a base para os estados mentais representacionais responsáveis pela ação intencional. Porém, “se os sistemas artificiais não são capazes de agência intencional, conforme interpretado pela teoria padrão, eles ainda podem ser capazes de algum tipo mais básico de agência.” (2019, p. 20. Tradução nossa)²², segundo pressupostos da própria teoria.

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

1.3 Agência e autonomia

Para dar continuidade a nossa análise, examinaremos ainda os conceitos de agência natural e agência artificial, e exploraremos o conceito de autonomia, característica considerada fundamental para que a agência possa ser pensada de maneira não trivial.

Notamos que a agência pode ser considerada em diversos contextos, porém a agência autônoma possui características bastante específicas. Buscamos assim as bases que nos permitirá caracterizar o que é um agente autônomo e, para tanto, faremos uso principalmente

²¹ In general, on relational autonomy accounts, autonomy is seen as an ideal by which we can measure how well an agent is able to negotiate his or her pursuit of goals and commitments, some of which may be self-chosen, and some the result of social and relational influences. Social and relational ties are examined in terms of their effect on an agent’s competency in this negotiation: some give strength, others create obstacles, and others are ambiguous. The primary focus of most relational autonomy accounts, however, tends to be less on procedure and more on changing the model of the autonomous self from an individualistic one to one embedded in a social context.

²² If artificial systems are not capable of intentional agency, as construed by the standard theory, they may still be capable of some more basic kind of agency.

dos escritos de Sarah Buss e Andrea Westlund (2018) e do já citado Donald Davidson (1963, 1971, 2002). Apresentaremos, ainda, outros autores que colaborarão com a nossa investigação.

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

Uma vez que nada e nem ninguém tem o poder de agir, exceto o próprio agente, apenas ele tem o direito de exercer esse poder, se tiver o direito de agir. Isso significa que, na medida em que alguém é um agente, ou seja, na medida em que é uma pessoa que age - ela está correta em considerar seus próprios compromissos com o ato, seus próprios julgamentos e decisões sobre como deve agir enquanto autoridade. De fato, se o agente fosse desafiar a autoridade, característica essencial de seus julgamentos e decisões, estas deixariam de ser suas próprias conclusões práticas. (Tradução nossa)²³

Entendemos que, para Buss e Westlund (2018), o poder de ação reside na parcela de autonomia necessária para que uma ação seja iniciada. Segundo ressalta Broens (2018), os agentes autônomos não são totalmente dependentes de seus contextos, existe uma margem para o direcionamento da ação, para a possibilidade de escolha, que é o momento no qual a autonomia se faz presente. Assim, levando em consideração as ideias apresentadas, é possível sustentar que o direito de ação seja justamente essa abertura contextual na qual o agente atua considerando fatores contextuais, mas não determinado por eles, alcançando resultados de suas próprias ações e experiências.

De acordo com Davidson (1963), a noção comum de causa que faz parte do discurso científico e o do senso comum é fundamental para a compreensão do que é agir de acordo com uma razão, ter certa intenção em agir, ser um agente e agir contra o próprio bom senso ou agir livremente.

Segundo o filósofo, o conceito de causa é o que unifica nossa imagem no universo. Ele considera as relações causais como sendo o “cimento” do universo, ou seja, o elemento aglutinador que unifica o universo de forma objetiva. Assim, diante do papel atribuído por Davidson às relações causais, entendemos que algumas de suas contribuições para a compreensão da natureza da ação são fundamentais para compreendermos o que, de fato, sustentaria uma ação deliberada.

²³ Since nothing and no one has the power to act except the agent herself, she alone is entitled to exercise this power, if she is entitled to act. This means that insofar as someone is an agent, i.e., insofar as she is one who acts—she is correct to regard her own commitments to acting, her own judgments and decisions about how she should act, as authoritative. Indeed, if she were to challenge the authority that is an essential feature of her judgments and decisions, then they would cease to be her own practical conclusions.

Além da contribuição dada por Davidson (1963), que possibilitou a construção de uma teoria padrão da ação, exposta na seção anterior, o filósofo (1971) propôs uma diferenciação entre eventos e ações na medida em que ações são consideradas eventos, porém nem todos os eventos constituem uma ação de fato. Desse modo, as ações, os atos e os feitos se diferenciam de acontecimentos não intencionais ou não deliberados, e essa diferenciação permitirá verificar quando uma ação é deliberada e, conseqüentemente, se o agente agiu por determinação própria.

Para ilustrar essas diferenças, Davidson (1971) relata uma manhã na qual ele foi acordado pelo som de alguém que praticava violino. Após ser acordado por um som, Davidson relata que tirou um breve cochilo, se levantou, se lavou e se barbeou, depois desceu as escadas e ao passar pelo corredor apagou a luz. Antes de pegar seu jornal, se serviu de café, tropeçou no tapete de sua sala e ainda derramou o café enquanto tateava pelo jornal.

O que Davidson quer nos dizer com seu relato é que, em algumas dessas circunstâncias, ele foi de fato o autor das mesmas, isso significa dizer que para que algumas dessas situações pudessem acontecer, ele teve de realizar algum tipo de esforço, enquanto que, em outras circunstâncias, como tropeçar no tapete ou ser acordado por um som inesperado, os eventos simplesmente aconteceram a ele enquanto deliberadamente realizava outras atividades. “Acordar é algo que eu fiz, talvez, mas não uma ação.” (DAVIDSON, 1971, p. 43. Tradução nossa)²⁴.

Segundo Davidson, diferenciar ações de meros acontecimentos é uma tarefa problemática, e é a presença da intencionalidade que poderia servir como diferenciador de ambos. Como vimos na primeira seção, para Davidson uma ação intencional é aquela que pode ser explicada através de alguma razão ou motivo, de forma que uma explicação racional indique a causa de uma ação. Assim, uma ação difere de um evento, ou de um mero acontecimento, se o agente puder explicar como ou porque executou determinada ação. Como diz o autor, “tropeçar em um tapete normalmente não é uma ação; mas seria se fosse feito intencionalmente.” (1971, p. 44. Tradução nossa)²⁵. Ainda assim, essa distinção não é suficiente para justificar que a intencionalidade implica agência, mas que a agência nem sempre implica intencionalidade.

Assim, derramar o café, afundar o Bismarck e insultar alguém são coisas que podem ou não ser feitas intencionalmente, mas mesmo quando não são intencionais, são normalmente ações. Se, por exemplo, eu intencionalmente derramar o conteúdo do meu copo, erroneamente pensando que é chá quando

²⁴ Waking up is something I did, perhaps, but not an action.

²⁵ Tripping over a rug is normally not an action; but it is if done intentionally.

é café, então derramar o café é algo que eu faço, é uma ação minha, embora eu não o faça intencionalmente. Por outro lado, se derramo o café porque alguém sacudiu a minha mão, não posso ser chamado de agente [dessa ação]. No entanto, embora eu possa apressar-me em acrescentar minha desculpa, não é incorreto, mesmo neste caso, dizer que derramei o café. Assim, devemos distinguir três situações em que é correto dizer que derramei o café: na primeira, faço-o intencionalmente; na segunda eu não faço isso intencionalmente, mas a ação é minha (achei que fosse chá); na terceira não é minha ação (alguém sacode a minha mão). (DAVIDSON, 1971, p. 45. Tradução nossa)²⁶

Neste sentido, em casos de agência, Davidson (1971) propõe que uma pessoa é autora de sua ação, ou seja, se o agente agiu intencionalmente, apenas se existir uma descrição sobre tal ação que faça com que a sentença que a expresse for verdadeira. Para o autor, essa formulação não é totalmente satisfatória, uma vez que quantifica entidades linguísticas (DAVIDSON, 1971).

Embora seja difícil encontrar concordância integral entre as diferentes teorias da agência, é possível perceber que estes estudos admitem algumas características essenciais da agência autônoma, tais como a possibilidade de justificar a própria ação e a capacidade de escolha, ou seja, a não coerção do agente para realizar a ação.

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

²⁶ Thus, spilling the coffee, sinking the Bismarck, and insulting someone are all things that may or may not be done intentionally, but even when not intentional, they are normally actions. If, for example, I intentionally spill the contents of my cup, mistakenly thinking it is tea when it is coffee, then spilling the coffee is something I do, it is an action of mine, though I do not do it intentionally. On the other hand, if I spill the coffee because you jiggle my hand, I cannot be called the agent. Yet while I may hasten to add my excuse, it is not incorrect, even in this case, to say I spilled the coffee. Thus, we must distinguish three situations in which it is correct to say I spilled the coffee: in the first, I do it intentionally; in the second I do not do it intentionally but it is my action (I thought it was tea); in the third it is not my action at all (you jiggle my hand).

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

(...) basta observar que se a fraqueza da vontade é um fenômeno genuíno, então os agentes humanos têm a capacidade de se governar de uma maneira que eles próprios consideram injustificada. Eles podem afirmar uma autoridade sobre si mesmos que desafia a autoridade de sua própria razão. (BUSS & WESTLUND, 2018, p. 15. Tradução nossa)²⁷

É possível compreender a noção de um agente “fraco de vontade” a partir de alguns exemplos nos quais os indivíduos se comportam de forma (pelo menos aparentemente) contrária a suas crenças, princípios e vontades.

Uma pessoa, por exemplo, que é compelida pelos seus colegas a fazer algo que inicialmente não está disposta a fazer pode ser considerada como sendo um agente com vontade fraca. Para ilustrar essa ideia podemos pensar facilmente em uma situação na qual todos os colegas se reúnem em um bar após o horário de trabalho. Quase todos na mesa têm seus copos cheios de cerveja, porém uma das pessoas resolve que aquele dia específico ela não está afim de consumir bebida alcoólica alguma. Em situações como essas, as pressões em forma de brincadeiras não são incomuns e não costumam cessar rapidamente, a ponto de fazer com que a pessoa que inicialmente não tinha a intenção nem a vontade de consumir bebida alcoólica, passe a consumi-la, ou que beba ‘pelo menos’ um copinho.

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

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

²⁷ (...) it suffices to note that if weakness of will is a genuine phenomenon, then human agents have the capacity to govern themselves in a way that they themselves take to be unjustified. They can assert an authority over themselves that challenges the authority of their very own reason.

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

Como foi dito, ainda existem muitas questões e controvérsias acerca da temática da agência, porém entendemos que as características apresentadas se mostram suficientes para a especificação de um **agente autônomo natural**. Sendo elas, em síntese, a possibilidade de autorregular-se sem coerções para planificar o futuro e/ou modificar tal planejamento.

Concluimos nossa investigação acerca da noção de agência autônoma em organismos naturais com as palavras de Buss e Westlund (2018, p. 17) que apontam que “[...] a única atitude da qual parece que nenhum agente pode ser alienado é o desejo de ter poder suficiente para determinar os próprios motivos - o desejo de ser um agente autônomo.” (Tradução nossa)²⁸. Mas será que sistemas artificiais poderiam ser considerados agentes nesse sentido forte, o qual inclui o desejo por autonomia?

1.4 Agência e autonomia em sistemas artificiais

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

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

²⁸ The only attitude from which it seems that no agent can be alienated is the desire to have sufficient power to determine one’s own motives—the desire to be a self-governing agent.

O *Google Brain* desenvolve IA que supostamente constrói IA melhor e mais rápido que os seres humanos. O *AlphaZero* pode se auto inicializar em quatro horas, ignorando completamente as regras do xadrez e o nível do campeão mundial. É impossível entender exatamente como *AlphaGo* conseguiu vencer o campeão humano de *Go World*. O aprendizado profundo e as chamadas "abordagens geradoras de redes contraditórias" permitem que as máquinas "ensinem" a si mesmas novas estratégias e busquem novas evidências para analisar. Nesse sentido, suas ações geralmente não são mais inteligíveis e não estão mais abertas ao escrutínio dos seres humanos. Esse é o caso porque, primeiro, é impossível estabelecer como eles alcançam seus resultados além dos algoritmos iniciais. Segundo seu desempenho é baseado nos dados que foram usados durante o processo de aprendizagem e que podem não estar mais disponíveis ou acessíveis. Assim, preconceitos e erros apresentados no passado ficam enraizados no sistema. (E.C., 2018, p. 6. Tradução e grifo nosso)²⁹

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

Em síntese, sistemas capazes de aprender tarefas sem a supervisão de um ser humano, ou seja, sem que sejam programados especificamente para realizar tarefas determinadas, podem ser considerados autônomos. Eles podem se manifestar como sistemas robóticos altamente tecnológicos ou como *softwares* inteligentes. “Muitos deles são liberados para o mundo sem supervisão e podem realizar coisas que não são previstas por seus projetistas ou proprietários humanos.” (E.C., 2018, p. 7. Tradução nossa)³⁰

Ressaltamos ainda que a declaração europeia sobre sistemas inteligentes (E.C., 2018, p. 9) considera que a autonomia está relacionada diretamente à dignidade humana e à

²⁹ Google Brain develops AI that allegedly builds AI better and faster than humans can. AlphaZero can bootstrap itself in four hours from completely ignorant about the rules of chess, to world champion level. It is impossible to understand how exactly AlphaGo managed to beat the human Go World champion. Deep learning and so-called ‘generative adversarial network approaches’ enable machines to ‘teach’ themselves new strategies and look for new evidence to analyze. In this sense, their actions are often no longer intelligible, and no longer open to scrutiny by humans. This is the case because, first, it is impossible to establish how they accomplish their results beyond the initial algorithms. Second, their performance is based on the data that have been used during the learning process and that may no longer be available or accessible. Thus, biases and errors that they have been presented with in the past become ingrained into the system.

³⁰ Many of them are released into the world unsupervised and may accomplish things which are not foreseen by their human designers or owners.

capacidade humana de autorregulação, a autonomia engloba ainda o direito de ser livre para escolher seus padrões, objetivos e propósitos na vida.

Os processos cognitivos que apoiam e facilitam essas capacidades estão entre os mais intimamente identificados com a dignidade das pessoas humanas, agência e atividade humana por excelência. Eles normalmente envolvem as características da autoconsciência e de autoria de acordo com razões e valores. (E.C., 2018, p. 9. Tradução nossa)³¹

Assim sendo, para a Comissão Europeia (2018), ainda que o conceito de autonomia tenha ganhado espaço nos debates científicos e no debate público para fazer referência a elevados graus de automação e de independência dos seres humanos em relação à operabilidade e à tomada de decisões, o termo **autonomia** não poderia ser empregado em outros sentidos e deveria ser utilizado apenas quando se trata de seres humanos.

A declaração (2018) afirma que a posição moral e a dignidade humana não devem ser transferidas aos sistemas artificiais, assim, os seres humanos deveriam ser sempre responsáveis por suas criações e permanecerem no controle das tecnologias.

Os seres humanos devem ser capazes de determinar quais valores são servidos pela tecnologia, o que é moralmente relevante, quais objetivos finais e concepções do bem são dignos de serem buscados. Isso não pode ser deixado para máquinas, não importa o quão poderosas elas sejam. (E.C., 2018, p. 10. Tradução nossa)³²

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

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

³¹ the cognitive processes that support and facilitate this are among the ones most closely identified with the dignity of human persons and human agency and activity par excellence. They typically entail the features of self-awareness, self-consciousness and self-authorship according to reasons and values.

³² Human beings ought to be able to determine which values are served by technology, what is morally relevant and which final goals and conceptions of the good are worthy to be pursued. This cannot be left to machines, no matter how powerful they are.

inteligente, é necessário refletir cuidadosamente como as decisões de um sistema artificial podem afetar vidas humanas.

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

Para dar continuidade e complementar nossa reflexão, apresentamos algumas considerações de Franklin e Graesser (1996), que propuseram definições e fizeram esclarecimentos sobre o que viria a ser a agência autônoma que vai ao encontro da definição proposta pela declaração europeia. Tais definições são bastante precisas e podem contribuir para que o objetivo proposto pela pesquisa seja alcançado³³.

Para Franklin e Graesser (1996) existem dois tipos de usos comuns para a palavra agente: o primeiro se refere a aquele que age, ou que pode agir, e o segundo uso da palavra se refere àquele que age, após permissão, no lugar de outro agente. “Como ‘aquele que age no lugar de’ age, o segundo uso requer o primeiro.” (1996, p. 25. Tradução nossa)³⁴

Assim, para ilustrar o primeiro uso do termo “agência” e dar exemplos que podem sustentar a definição que está sendo construída, os autores apresentam três tipos de agentes: os seres humanos e os animais, os *software agents* que “vivem”³⁵ em sistemas operacionais de computadores, banco de dados e redes, e por fim agentes artificiais (*artificial life agents*) que “vivem” em ambiente artificiais na tela de um computador ou em sua memória (FRANKLIN e GRAESSER, 1996, p. 25). E os requisitos que constituem a “essência” de um agente autônomo são:

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

³³ A declaração da Comissão Europeia (2018) trata de *softwares*, programas e sistemas inteligentes atualmente elaborados ou já tecnicamente concebíveis, enquanto o trabalho de Franklin e Graesser, publicado em 1996, contribui com a questão conceitual envolvida na problemática que envolve a agência autônoma de sistemas artificiais.

³⁴ Since "one who acts in place of " acts, the second usage requires the first.

³⁵ Destaque do autor.

continuamente ao longo de algum período de tempo. (FRANKLIN e GRAESSER, 1996, p. 25. Tradução nossa)³⁶

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

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

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

De forma resumida, é possível afirmar que para um sistema ser considerado um agente ele deve estar situado em um ambiente e ser dependente das condições que determinado ambiente fornece para que ele possa atuar nesse meio. Ademais, um agente pode ser considerado autônomo quando é capaz de escolher suas ações de forma independente enquanto cumpre sua agenda: “Agentes autônomos estão situados em algum ambiente. Altere o ambiente e talvez não tenhamos mais um agente. Um robô com sensores apenas visuais em um ambiente sem luz não é um agente.” (FRANKLIN e GRAESSER, 1996, p. 26. Tradução nossa)³⁷

Neste sentido, consideramos que, embora a Declaração de 2018 não concorde em classificar agentes artificiais como agentes autônomos no sentido forte do termo, ou seja, agentes conscientes de sua capacidade de ação e de escolha, se um agente artificial, ainda que a cumprir ordens, exiba comportamentos e *outputs* que não foram previstos, se ele é capaz de

³⁶ Each is situated in, and is a part on some environment. Each senses its environment and act autonomously upon it. No other entity is required to feed it input, or to interpret and use its output. Each acts in pursuit of its own agenda, whether satisfying evolved drives as in humans and animals, or pursuing goals designed in by some other agent, as in software agents. (Artificial life agents may be of either variety.) Each acts so that its current actions may affect its later sensing, that is its actions affect its environment. Finally, each acts continually over some period of time.

³⁷ Autonomous agents are situated in some environment. Change the environment and we may no longer have an agent. A robot with only visual sensors in an environment without light is not an agent.

escolher entre opções de acordo com o ambiente e contexto no qual está inserido, indo além de sua programação inicial e aprendendo com seu ambiente, esse agente pode ser considerado como agente portador de algum grau de autonomia.

Concluimos assim o primeiro capítulo após caracterizar as noções de agência, ação e autonomia, e verificar que sistemas artificiais, em determinados contextos, cumprem com as condições necessárias para serem considerados como sistemas portadores tanto de agência quanto de autonomia.

Dando continuidade à nossa investigação que tem como objetivo verificar se é possível atribuir responsabilidade a sistemas artificiais, no segundo capítulo faremos uma caracterização da conduta moralmente significativa, bem como da conduta moral referente ao uso de artefatos tecnológicos. Investigaremos ainda algumas noções do conceito de “responsabilidade” e a possibilidade da atribuição de coletiva. Por fim, investigaremos em quais condições a atribuição de responsabilidade a sistemas artificiais seria adequada.

CAPÍTULO 2

CONDUTA MORAL E TIPOS DE

RESPONSABILIDADE

APRESENTAÇÃO

A questão da agência moral vem sendo discutida pelos filósofos das mais diferentes tradições há muito tempo, sem que tenham chegado a um acordo definitivo sobre a melhor forma de defini-la. De modo geral, filósofos da tradição ocidental, como Anscombe (1975) e Davidson (1963), defenderam que apenas seres humanos podem ser legitimamente considerados agentes morais, uma vez que somente seres humanos atuariam considerando valores significativos e teriam a capacidade de adotá-los, contestá-los ou ignorá-los, isto é, só seres humanos seriam agentes intencionais responsabilizáveis pelas consequências de suas ações. Apenas seres humanos seriam agentes intencionais porque, ainda segundo a tradição filosófica ocidental, só eles dominariam as capacidades cognitivas consideradas necessárias para o exercício da agência, dentre as quais se destacam: (1) a capacidade de contribuir para a construção de uma cultura que, ao contrário do ambiente natural, estaria carregada de significado e (2) o uso das línguas naturais para a produção de conceitos e teorias, inclusive teorias morais.

Nas perspectivas antropocêntricas e representacionistas, a agência moral estaria condicionada: à (1') capacidade de auto-referenciação (o uso da primeira pessoa *eu*) que permite a expressão de (2') um desejo ou uma crença (*eu desejo x* ou *eu acredito que x*) e, por fim, de (3') desejos e crenças associados a valores morais (*eu desejo o bem* ou *eu acredito que a ação x é boa*) (4') *livremente* adotados e por cujas consequências (5') os agentes morais podem ser legitimamente *responsabilizados*.

Consideramos que esta concepção de agência é antropocêntrica principalmente porque utiliza, sem justificativa, o tipo humano de agência moral como padrão universal dentro do qual eventuais candidatos a serem reconhecidos como agentes morais precisariam se encaixar, como se a excelência da agência moral humana fosse autoevidente. Atualmente, porém, estudos interdisciplinares têm evidenciado que o tipo humano de agência moral não é exclusivo, tendo sido descobertos evidências consideráveis de ricas vidas morais em animais não humanos.

Com o intuito de esclarecer a noção de responsabilidade, inicialmente introduziremos algumas diferenciações do uso do termo “responsabilidade” elaboradas por Hans Jonas (2006). Por fim, apresentaremos uma caracterização de responsabilidade coletiva derivada para verificar se a questão inicial, que indaga se agentes artificiais autônomos podem ser legitimamente incluídos no grupo de agentes responsabilizáveis, pode ser satisfeita em algum sentido que não seja trivial.

2.1 Atribuição de responsabilidade

No que diz respeito à atribuição de responsabilidade, alguns estudos sobre a conduta moral sugerem que um agente pode ser responsabilizado apenas se suas ações ou disposições forem voluntárias, isto é, foram realizadas sem coerção. No primeiro capítulo tratamos sobre agência e autonomia e apresentamos algumas características fundamentais que ilustram nosso entendimento de uma ação livre e sem coerção.

Introduzimos agora o filósofo Hans Jonas para tratarmos sobre o tema da responsabilidade e entendermos como a responsabilidade nem sempre dirá respeito a um sentimento moral.

Jonas, em sua obra *O Princípio Responsabilidade - Ensaio de uma ética para a civilização tecnológica* (2006), tratou da importância da noção de responsabilidade no contexto das tecnologias, especialmente as de uso bélico, argumentando que, ainda que o fenômeno da responsabilidade não represente algo novo para os estudos da moralidade, as teorias éticas concederam pouca atenção a tal fenômeno. (JONAS, 2006, p. 22).

O filósofo, fazendo referência ao início das civilizações, quando os seres humanos estavam a construir suas cidades e firmar suas identidades culturais, aponta que: “[...] tanto o conhecimento quanto o poder eram por demais limitados para incluir o futuro mais distante em suas previsões e o globo terrestre na consciência da própria causalidade” (JONAS, 2006, p. 22). Para Jonas, as teorias éticas então em uso (muitas vezes vinculadas a sentimentos religiosos), se preocupavam com a ação no âmbito do aqui e do agora, ou seja, no âmbito temporal em que ocorre a ação, uma vez que não havia a necessidade de se preocuparem com as consequências das ações a longo prazo, na medida em que as necessidades e criações dos seres humanos ainda não eram capazes e suficientes para modificar a natureza de forma irreversível.

De acordo com Jonas, a capacidade humana de produzir tecnologias, cujo impacto sobre o meio ambiente seja irreversível ou provoque danos de longuíssima duração, tem aumentado a tal ponto que as tecnologias necessitam, cada vez mais, serem discutidas de uma perspectiva ética quanto a suas implicações, sendo que a própria sobrevivência da espécie pode vir a ser ameaçada pelo desenvolvimento de novas tecnologias.

Segundo Jonas (2006) as primeiras criações tecnológicas humanas visavam suprir necessidades com fins precisos e delimitados (pensemos em ferramentas, por exemplo). Já a técnica contemporânea, ressalta Jonas:

[...] a *techne* transformou-se em um infinito impulso da espécie para adiante, seu empreendimento mais significativo. [...] a tecnologia assume um significado ético por causa do lugar central que ela agora ocupa subjetivamente nos fins da vida humana. Sua criação cumulativa, isto é, o meio ambiente artificial em expansão, reforça, por um contínuo efeito retroativo, os poderes especiais por ela produzidos: aquilo que já foi feito exige o emprego inventivo e incessante daqueles mesmos poderes para manter-se e desenvolver-se, recompensando-o com um sucesso ainda maior - o que contribui para o aumento de suas ambições. (JONAS, 2006, p. 43)

Entendemos assim que, no momento em que a humanidade passou a buscar controlar a natureza (com sucesso limitado) e a manipulá-la de maneira drástica, desenvolvendo tecnologias que ameaçam a sua sobrevivência e a de outras espécies, tornou-se necessário pensar em uma nova ética, a ética da responsabilidade. Nas palavras do filósofo, essa nova ética da responsabilidade deve ter como princípio um imperativo adequado ao novo tipo de agir humano, o qual pode ser expresso de várias formas:

“Atue de modo a que os efeitos de seus atos sejam compatíveis com a permanência de uma autêntica vida humana sobre a Terra”; ou, expresso negativamente: “Atue de modo a que os efeitos de seus atos não sejam destrutivos para a possibilidade futura de uma tal vida”; ou simplesmente: “não ponha em perigo as condições necessárias para a conservação indefinida da humanidade sobre a Terra”; ou, em um uso novamente positivo: “Inclua na tua escola presente a futura integridade do ser humano como um dos objetos do seu querer.” (JONAS, 2006, p. 47-48)

No que se refere às questões sobre responsabilidade, Hans Jonas (2006) afirma que só será possível tratar sobre uma responsabilização quando o agente for causa ativa da ação, ou seja, quando o agente tiver ligação direta com a ação praticada.

O poder causal é condição da responsabilidade. O agente deve responder por seus atos: ele é responsável por suas consequências e responderá por ela, se for o caso. Em primeira instância, isso deve ser compreendido do ponto de vista legal, não moral. Os danos causados devem ser reparados, ainda que a causa não tenha sido um ato mau e suas consequências não tenham sido nem previstas, nem desejadas. Basta que eu tenha sido a causa ativa. Mas isso somente se houver um nexo causal estreito com a ação, de maneira que a imputação seja evidente e suas consequências não se percam no imprevisível. (JONAS, 2006, p. 165).

Uma das novidades que Jonas traz diz respeito à diferenciação que sugere entre duas maneiras diferentes de responsabilizar um agente, pois a compensação legal se diferenciaria da punição. Entende Jonas (2006) que, na compensação de um ato, o agente deve ser reconhecido como responsável e deve, como dito acima, reparar seus atos, mas sem necessariamente carregar o estigma e a culpa daquele que prejudica outrem com suas ações.

No entanto, diferentemente da compensação legal, a punição tem caráter moral e “qualifica o ato causal como moralmente culpabilizável.” (JONAS, 2006, p. 165). Segundo o

filósofo, ser julgado culpado por um ato é diferente de dever uma compensação ou reparação a outrem devido aos seus atos. E, após investigar o ato moralmente prejudicial tentando descobrir qual foi a sua motivação, a punição a ser aplicada, ou seja, a devida responsabilização do agente, não terá como objetivo a compensação por danos ou injustiças, mas sim o restabelecimento da ordem moral que foi perturbada.

Nos termos da lei, toda responsabilidade moral pode gerar responsabilidade legal, mas nem toda responsabilidade legal gera responsabilidade moral. Uma pessoa que, por sua ação, prejudica o exercício de um direito de alguém, incorre necessariamente em uma ilegalidade, mas não necessariamente comete um crime ou é moralmente responsável, por isso uma pena não lhe é aplicável, mas deve ainda assim compensar o prejuízo causado.

A responsabilização legal implica respeitar a lei, implica respeitar o direito de alguém conforme a legislação em vigor. A responsabilização moral implica o cumprimento de dever, agindo ou deixando de agir, compreendendo que a moral tem como um de seus princípios centrais o respeito à dignidade de uma pessoa.

A diferença apontada entre a responsabilidade legal e a moral reflete-se na diferença entre o direito civil e o direito penal, em cujas evoluções divergentes dissociaram-se os conceitos inicialmente mesclados de compensação (como responsabilidade legal) e pena (pela culpa). Ambos têm em comum que a “responsabilidade” se refere a atos realizados, e que a responsabilização efetiva do autor é feita desde o exterior. (JONAS, 2006, p. 166)

Para o filósofo, na inexistência de um dever positivo, ou seja, na falta da necessidade de uma ação que deve ser, de fato, executada, evitar agir pode ser prudente na medida em que quanto menos ações houver, menos responsabilizações haverá.

(...) entendida assim, a “responsabilidade” não fixa fins, mas é a imposição inteiramente formal de todo agir causal entre seres humanos, dos quais se pode exigir uma prestação de contas. Assim, ela é condição da moral, mas não a própria moral. O sentimento que caracteriza a responsabilidade (...) é de fato moral (disposição de assumir seus atos), mas em sua formalidade pura não é capaz de fornecer o princípio efetivo para a teoria ética, que em primeira e última instância tem a ver com a apresentação, reconhecimento e motivação de finalidades positivas para o *bonum humanum*. (JONAS, 2006, p. 166, grifo nosso)

Jonas tratou ainda de diversos tipos de responsabilidade, tais como: responsabilidade parental, responsabilidade política, responsabilidade natural e responsabilidade contratual, porém não nos debruçaremos sobre esses tipos, uma vez que sua distinção não afeta o tipo específico de responsabilidade do qual tratamos: é responsabilizável todo agente cujo ato implique um dano a outrem, uma consequência imoral ou ilegal; esse agente é passível de penalização e da obrigação de compensar os danos que causou.

Porém, quando pensamos a responsabilidade em um futuro ou contexto incerto, como é o caso dos efeitos que podem ser causados por tecnologias autônomas e inteligentes, há outro tipo de responsabilidade descrita por Jonas, responsabilidade do “dever do poder” fazer algo.

Jonas (2006) esclarece que esse tipo de responsabilidade não está propriamente ligado ao reconhecimento da culpa ou da necessidade de reparação de danos provocados pela ação do agente. Essa responsabilidade ocorre, por exemplo, quando alguém reconhece poder fazer algo e assume esse poder como um dever, por exemplo, quando um agente reconhece sua responsabilidade de agir solidariamente ou sua responsabilidade por eventuais efeitos que uma tecnologia que idealizou pode vir a causar. Desse modo, a possibilidade de imputação/reconhecimento de responsabilidade por atos que podem vir a ser causados, pode auxiliar-nos a pensar a questão da atribuição de responsabilidade nos novos contextos tecnológicos, além da possibilidade da atribuição de responsabilidade coletiva, que veremos mais adiante.

Em síntese, Jonas (2006) apresenta a noção de responsabilidade, a do “dever do poder” que não mais se refere àquilo que foi feito e suas eventuais consequências, mas sim a uma determinação do que é necessário fazer no futuro. Trata-se de uma concepção:

[...] em virtude da qual eu me sinto responsável, em primeiro lugar, não por minha conduta e suas consequências, mas pelo objeto que reivindica meu agir. Responsabilidade, por exemplo, pelo bem-estar de outros, que considera determinadas ações não só do ponto de vista da sua aceitação moral, mas se obriga a atos que não têm nenhum outro objetivo. (JONAS, 2006, p. 167).

Para exemplificar situações nas quais o agente é detentor de um poder, podemos pensar em uma viagem de avião ou até mesmo em um parque de diversões: no primeiro caso, o piloto se coloca como responsável pela segurança e vida de todos os passageiros e a tripulação, na medida em que deve transportá-los de maneira segura até seu destino. É fato que, em uma companhia aérea, existem cargos superiores ainda aos de um piloto e seus detentores são passíveis de responsabilização por eventuais falhas, porém, no momento em que o piloto se propõe a entrar no avião e conduzir os passageiros a seu destino, a responsabilidade pelo bem estar dos passageiros é a ele atribuída, e o sentimento moral do dever relativo a essa responsabilidade é sentido pelo piloto na medida em que apenas ele pode julgar se está ou não apto a pilotar um avião com perícia. No segundo exemplo, no parque de diversões, o dono do parque é responsável pela segurança das pessoas que desfrutam de seus brinquedos, assim ele é também responsável pelos profissionais que são contratados para

fazer a manutenção dos aparelhos, por aqueles que controlam quem pode ou não entrar nos brinquedos, pelas empresas de alimentos dentro de seu parque, etc.

É neste sentido que Jonas propõe o que significa agir de maneira irresponsável em casos nos quais um poder é deliberadamente concedido a um agente.

Por circunstâncias ou por convenção, encontram-se sob meus cuidados o bem-estar, o interesse e o destino de outros, ou seja, o controle que tenho sobre eles inclui, igualmente, a minha obrigação para com eles. O exercício do poder sem a observação do dever é, então, “irresponsável”, ou seja, representa uma quebra da relação de confiança presente na responsabilidade. (JONAS, 2006, p. 168, grifo nosso)

Por fim, relembro as colocações de Jonas (2006) que se referem (1) à técnica moderna como um caminho para que a humanidade atinja seu fim e não mais apenas como a um modo de buscar sanar necessidades imediatas, (2) à vocação humana de buscar um progresso ou avanço contínuo (p. 43), queremos investigar neste trabalho os possíveis e prováveis riscos que o futuro reserva quanto ao uso de tecnologias capazes de agir com algum grau de autonomia e a possibilidade (ou não) de incluir agentes artificiais na rede de responsabilidade descrita por Jonas. Antes de tratar dessa questão, porém, trataremos de um novo tipo de responsabilidade, a coletiva, que, entendemos, pode nos auxiliar em nossa análise sobre as implicações da autonomia de agentes artificiais.

2.2 Responsabilidade coletiva e ações em conjunto

No que diz respeito à responsabilidade coletiva, estudos sugerem que em toda ação coletiva existe um núcleo de membros responsáveis pela ideia a ser executada (muitos grupos sociais, senão a maioria, possuem uma estrutura hierárquica de organização). Porém, ainda que seja possível identificar esse núcleo, todos os agentes envolvidos na execução da tarefa podem ser responsabilizados, em algum sentido, pelas ações realizadas pelo grupo.

De acordo com Marion Smiley (2017), responsabilidade **coletiva** diz respeito à atribuição de responsabilidade a uma entidade coletiva, como, por exemplo, corporações, Estados, sociedades, empresas, clubes, etc. Assim, quando membros de uma entidade coletiva executam, autorizam a execução ou simplesmente se omitem de evitar uma ação prejudicial, a responsabilidade por esta ação será atribuída à entidade e não a seus membros.

A responsabilidade **compartilhada**, por sua vez, diz respeito à responsabilidade atribuída a membros de um grupo que causaram danos a partir de suas ações.

A responsabilidade coletiva está associada a um agente moral unificado. A responsabilidade compartilhada está associada a agentes morais individuais que contribuem para prejudicar os membros de um grupo diretamente por

meio de suas próprias ações ou indiretamente por meio de sua participação no grupo. (SMILEY, 2017, p. 26. Tradução nossa)³⁸

Seumas Miller, em *Collective Responsibility and Information and Communication Technology* (2008), examina a noção de responsabilidade coletiva nos domínios da comunicação e recuperação de conhecimento através das Tecnologias de Informação e de Comunicação (TIC) e nos termos da teoria dos fins coletivos (p. 226).

Para realizar sua análise, Miller (2008) utilizou conceitos que, segundo ele, são de extrema importância para o estudo da responsabilidade nos termos das TIC, tais como ação conjunta, ação conjunta epistêmica, asserções, convenções ou procedimentos conjuntos, normas sociais, leis, leis morais, fins coletivos, informações e dados. Para alcançar os propósitos da presente investigação, apresentaremos alguns destes conceitos de forma mais detalhada, enquanto outros serão apenas pincelados de forma a não prejudicar o entendimento do texto. Este trabalho nos permitirá compreender os fatores essenciais que possibilitam a caracterização de um agente responsabilizável, seja ele natural ou artificial.

Assim, o primeiro conceito importante a ser explicitado na teoria de Miller é o de “ação conjunta” (*joint action*). De acordo com o autor, as ações conjuntas “[...] envolvem um número de agentes realizando ações interdependentes a fim de realizar algum objetivo comum” (MILLER, 2008, p. 227. Tradução nossa)³⁹. Como exemplo de ações conjuntas, podemos pensar em duas pessoas dançando, colegas se reunindo a fim de organizar um evento, trabalhadores construindo uma casa, etc.

Para Miller, ações conjuntas devem ser diferenciadas das ações individuais e das ações de órgãos corporativos. Ações individuais constituem ações nas quais o agente é o único protagonista da ação, por exemplo, ao montar sozinho um quebra cabeça, preparar uma refeição, ler um livro, caminhar, etc. Por sua vez, exemplos de ações corporativas podem ser “[...] uma nação que declara guerra ou um governo que toma uma ação legal contra uma empresa pública [...]” (MILLER, 2008, p. 227. Tradução nossa)⁴⁰.

Segundo o autor, o conceito de “ação conjunta” pode ainda ser entendido de maneira restrita ou de forma ampla. No primeiro caso, a ação conjunta é interpretada como sendo uma ação conjunta básica com dois agentes executando simultaneamente ações individuais com

³⁸ Collective responsibility is associated with a single, unified, moral agent. Shared responsibility is associated with individual moral agents who contribute to harm as members of a group either directly through their own actions or indirectly through their membership in the group.

³⁹ Joint actions are actions involving a number of agents performing interdependent actions in order to realize some common goal.

⁴⁰ A nation declaring war or a government taking legal action against a public company are instances of corporate action.

vista a fim coletivo imediato. “Um exemplo de uma ação individual básica é levantar o braço; um exemplo de uma ação conjunta básica são duas pessoas cumprimentando-se com um aperto de mãos.” (MILLER, 2008, p. 228. Tradução nossa)⁴¹.

No segundo caso uma forma mais ampla de analisar uma ação conjunta pode ser explicitada como segue:

Muitas ações conjuntas envolvem intenções e fins dirigidos a resultados fora do horizonte temporal da experiência imediata dos agentes, por exemplo, duas pessoas conversando durante duas, três horas ou pessoas decidindo construir um muro para o jardim durante as férias de verão. Outras envolvem intenções e fins dirigidos a resultados que existirão fora do horizonte espacial da experiência imediata dos agentes e envolvem outros instrumentos além dos corpos dos agentes. (MILLER, 2008, p. 228. Tradução nossa)⁴²

A interpretação mais ampla de ação conjunta é de extrema relevância para os estudos das tecnologias informacionais, uma vez que a revolução digital e o desenvolvimento das TIC fizeram com que esse tipo de ação se tornasse um parâmetro para relações sociais, principalmente no que diz respeito a atividades profissionais cotidianas.

Segundo Miller, novas formas de trabalho conjunto surgiram e as TIC ampliaram de forma considerável o leque de ações conjuntas possíveis. Enfatiza Miller (2008, p. 228) que: “Trabalhadores localizados em diferentes partes do mundo podem trabalhar durante longos períodos de tempo como uma equipe trabalhando em um projeto conjunto com objetivos comuns.” (Tradução nossa)⁴³. No que diz respeito aos objetivos comuns, também entendidos como *fins coletivos*, Miller aponta que estes são uma construção fora da noção prévia de um fim individual, ou seja, quando um fim é coletivo, mais de um agente está de acordo com este fim e trabalha para efetivá-lo e, assim sendo, todas as ações dos agentes envolvidos em determinado fim coletivo são fundamentais.

Miller destaca uma categoria particular das ações conjuntas envolvidas nas TIC: a asserção, ou seja, afirmações e declarações, que são os atos de fala, ou como afirma Miller (2008, p. 230) “[...] tipos de ações cognitivas ou epistêmicas, a saber, as chamadas atitudes ou ações que apontam para a verdade.” (Tradução nossa)⁴⁴. Assim, Miller passa a tratar dos atos

⁴¹ An example of a basic individual action is raising one’s arm; an example of a basic joint action is two people greeting one another by shaking hands.

⁴² Many of these involve intentions and ends directed to outcomes outside the temporal horizon of the immediate experience of the agents, for example, two people engaging in a two-hour long conversation or three people deciding to build a garden wall over the summer break. Others involve intentions and ends directed to outcomes that will exist outside the spatial horizon of the immediate experience of the agents and involve instruments other than the agents’ bodies.

⁴³ Workers located in different parts of the world can over lengthy time periods function as a team working on a joint project with common goals.

⁴⁴ [...] kinds of cognitive or epistemic actions, namely, so-called truth-aiming attitudes or actions.

de fala que, para ele, são principalmente *uma espécie de ação conjunta* que estão presentes em cada nível do processo informacional.

A prática dessas asserções que envolvem informação possui, segundo Miller, três características interligadas: (1) as asserções têm um objetivo ou fim comunicativo; levando em consideração que os participantes dos atos de fala têm em vista a verdade, ou que tentam ao menos evitar a falsidade, (2) são atos pelo menos restritos por considerações de veracidade, e por fim, (3) aqueles que proferem o discurso, não apenas visam a verdade, mas devem mostrar que buscam por ela. Neste sentido, é através das asserções que a informação é passada de um interlocutor a outro, e para o autor não existe informação falsa, (tese que nós assumimos neste trabalho, pois conteúdos falsos seriam desinformativos, não informativos), como esclarece a seguir:

Para o que se segue assumo que a informação é: (i) o conteúdo proposicional de atitudes que visam a verdade, por exemplo, crenças e ações que visam a verdade, por exemplo, asserções; e (ii) conteúdo proposicional verdadeiro. Consequentemente, afirmações falsas não transmitem informação, no meu sentido do termo; então não existe informação falsa. Além disso, proposições em abstrato, por exemplo, proposições que ninguém nunca, ou jamais, expressará ou pensa, não constituem informação. Proposições no sentido de conteúdo que não é, como tal, direcionador da verdade e, portanto, não passível de avaliação como verdadeiro ou falso, por exemplo, conteúdo da forma 'quer o gato esteja ou não no tatame' ou 'o gato estar no tatame' (diferentemente de 'o gato está no tatame') também não constitui informação. Finalmente, a informação, assim definida, é adequada para uso como elemento (s) na inferência, por exemplo, como a premissa em um argumento dedutivo. Para a assim chamada "informação" que é comunicada, armazenada ou recuperada, mas que é falsa ou não-proposicional no sentido acima, usarei o termo "dado". (MILLER, 2008, p. 230-231. Tradução nossa)⁴⁵

Após defender que os atos de fala são ações conjuntas que envolvem pelo menos dois agentes que, em princípio, devem estar preocupados com a verdade, Miller (2008) passa a considerar os atos de fala expressos através da escrita e, segundo o autor, a partir de um único ato de fala é possível atingir muitos 'ouvintes'. De fato, a linguagem escrita permite exatamente que um único agente que profira algum tipo de asserção atinja uma audiência em

⁴⁵ I will assume in what follows that information is: (i) the propositional content of truth-aiming attitudes, for example, beliefs, and of truth-aiming actions, for example, assertions; and (ii) true propositional content. Accordingly, false assertions do not convey information, in my sense of the term; so there is no such thing as false information. Moreover, propositions in the abstract, for example, propositions that no-one has ever, or will ever, express or think of, do not constitute information. Propositions in the sense of content that is not, as such, truth-aiming, and, therefore, not assessable as true or false, for example, content of the form 'whether or not the cat is on the mat' or 'the cat's being on the mat' (as distinct from 'the cat is on the mat') do not constitute information either. Finally, information, thus defined, is suitable for use as an element(s) in inference-making, for example, as the premise in a deductive argument. For so-called information that is communicated, stored, or retrieved, but which is false or nonpropositional in the above sense, I will use the term 'data'.

qualquer lugar que ela esteja, e por sua vez, as TIC permitem que essas asserções sejam proferidas mais rapidamente e para um maior número de espectadores.

Miller (2008) se propôs investigar a atribuição de responsabilidade nos domínios da comunicação, do armazenamento e do compartilhamento da informação através das TIC, e não a atribuição de responsabilidade nos domínios das tecnologias autônomas e inteligentes. Porém, entendemos que as TIC são uma parte importante dessas tecnologias, uma vez que sistemas inteligentes operam através de dados recebidos e armazenados em seus componentes.

Miller apresentou ainda conceitos fundamentais que auxiliam a compreensão dos comportamentos dos usuários das TIC, o funcionamento das máquinas, além da relação e a diferença entre leis e conduta moral. Assim, na próxima seção daremos continuidade à nossa investigação abordando o significado da ação moral e responsável com relação às TIC e às tecnologias inteligentes.

2.3 Conduta moral responsável e novas tecnologias de informação

Após diferenciar algumas noções em torno a vários sentidos do conceito de responsabilidade, e apontar que se uma pessoa ou grupos de pessoas são responsáveis por alguma ação, em sentido amplo, significa dizer que esta pessoa ou grupo de pessoas teve razões e intenções antes de executar ou não uma ação, Miller (2008, p. 240) afirma:

Uma ação pode ser moralmente significativa de várias maneiras. A ação poderia ser intrinsecamente moralmente errada, como no caso de uma violação de direitos. Ou a ação pode ter um significado moral em virtude do fim para o qual foi realizada ou do resultado que realmente teve. (Tradução nossa)⁴⁶

Neste sentido, a responsabilidade moral coletiva é entendida como uma responsabilidade conjunta, embora em uma ação conjunta cada agente tenha a sua responsabilidade individual estabelecida, assim como cada agente tem de executar uma ação determinada para que um plano conjunto se efetive, ainda assim os indivíduos estão condicionados às responsabilidades alheias e são interdependentes no sentido da responsabilidade moral.

O autor diferencia ainda a responsabilidade moral pelos estados epistêmicos da responsabilidade moral pelas ações. De acordo com as suas palavras, os estados epistêmicos são, ou deveriam ser, dependentes de processo racional, tal como a inferência dedutiva, uma

⁴⁶ An action can be morally significant in a number of ways. The action could be intrinsically morally wrong, as in the case of a rights violation. Or the action might have moral significance by virtue of the end that it was performed to serve or the outcome that it actually had.

vez que esta compele através de suas evidências, não deixando espaço, ou deixando pouco espaço, para a vontade. Ou seja, os estados epistêmicos moralmente significativos são diferentes de uma ação pela ação, uma vez que trabalham em conjunto com a racionalidade e com evidências que a suportem. Assim sendo, o autor sugere ainda que essa diferença resulta num contraste com a tomada de decisão prática (Miller, p. 246).

Neste sentido, e nos termos das TIC, aponta Miller (2008) que a responsabilidade moral poderia ser referente à comunicação, armazenamento e recuperação de informações moralmente desejáveis.

Já vimos que as asserções são uma espécie de ação conjunta e, no caso de asserções moralmente significativas, os oradores e o público podem, em consequência dessas asserções, razoavelmente ser responsabilizados coletivamente pelas opiniões em que acreditam. Na medida em que as TIC envolvem a comunicação, o armazenamento e a recuperação de asserções moralmente significativas, os usuários que são propagadores e público delas podem, do mesmo modo, ser responsabilizados coletivamente pelas crenças resultantes das afirmações mediadas por computador. (MILLER, 2008, p. 247. Tradução nossa)⁴⁷.

Entendemos assim que, quando se trata do compartilhamento de informação envolvido nas TIC, os usuários que compartilham e recebem as informações são responsáveis pela verificação de verdade de seus conteúdos.

Embora não seja o foco de nossa investigação, sabe-se que uma das problemáticas centrais do compartilhamento massivo de informação através das novas tecnologias de comunicação é a produção proposital, compartilhamento e disseminação (intencional ou ingênua) de notícias falsas (*fake news*), ou seja, notícias que geralmente têm como objetivo enganar e dissuadir os leitores. A confusão causada pela disseminação desenfreada de *fake news* tem muitas vezes como consequência episódios de violência e danos de diversos tipos, tanto físicos quanto morais de pessoas ou grupos alvos desse tipo de notícia, em geral difamatória e carregada de opiniões preconcebidas. Assim, é fundamental que tanto quem recebe, quanto quem compartilha tais notícias se responsabilize pelo conteúdo das mesmas e os eventuais danos que possam causar.

Miller (2008) conclui seu trabalho indicando dois pontos gerais que relacionam a responsabilidade moral e os sistemas das TIC. O primeiro diz respeito a uma distinção que

⁴⁷ We have already seen that assertions are a species of joint action, and in the case of morally significant assertions, speakers and audiences can reasonably be held collectively morally responsible for the comings to believe consequent upon those assertions. In so far as ICT involves the communication, storage, and retrieval of morally significant assertions, users of ICT who are speakers and audiences can likewise reasonably be held collectively morally responsible for the comings to believe consequent upon those computer-mediated assertions.

deve, segundo ele, ser feita entre a aplicação de procedimentos mecânicos, tanto por seres humanos ou por computadores, e entre a interpretação de princípios morais e leis.

Segundo Miller (2008), os computadores e outros sistemas especialistas, como aqueles atualmente utilizados em processos judiciais, não trabalham com significados e com interpretação, ou seja, esses sistemas artificiais não trabalhariam com o plano semântico das asserções, trabalham apenas com a sintaxe. Neste sentido, sistemas artificiais seriam apenas um meio para a comunicação (e facilitação do trabalho judiciário, por exemplo), sendo que eles não são capazes de julgar o que é certo ou errado, ou o que é verdadeiro ou falso.

O segundo ponto geral afirma que a conformidade com convenções e leis de sistemas mecânicos são significativamente diferentes da conformidade com princípios e fins de agentes morais. Para Miller (2008, p.249), é um erro supor que aquilo que é exigido ou permitido pela moralidade em certas situações deve ser a mesma coisa que a lei exige ou permite na mesma situação e menos ainda quando se trata de um procedimento mecânico.

Assim, em contraste às leis e aos procedimentos mecânicos, especialmente aos conduzidos por sistemas artificiais especialistas, a moralidade é um instrumento poderoso (*sharp instrument*), e deveria, segundo Miller (2008), ser aplicado a situações determinadas levando em consideração todas as suas particularidades. O problema identificado quando o assunto são leis e decisões morais é que, levando em consideração a particularidade inerente da tomada de decisão moral, ela não pode ser capturada totalmente pelos sistemas jurídicos e seus processos de interpretação e menos ainda pelos sistemas artificiais especialistas e seus processos predeterminados de aplicação mecânica de procedimentos.

Ao tratar sobre sistemas inteligentes portadores de agência que são idealizados por agentes capazes de interpretar significados morais, Miller (2008) entende que, estes sistemas e seus *outputs* são de responsabilidade de seus idealizadores, na mesma medida em que aqueles que criam e compartilham informações são responsáveis pelo seu conteúdo. Os criadores, elaboradores, desenvolvedores de sistemas podem ser identificados como um grupo que engloba tanto as pessoas que encomendam os sistemas, quanto aquelas que os constroem, os programam e que os testam.

Mas, tendo em vista que os sistemas tecnológicos estão sendo atualmente elaborados com a capacidade de se adaptarem ao seu ambiente, não precisando sempre de ordens e comandos para executar suas tarefas e agendas. Neste sentido, talvez, contrariando as teses de Miller (2008), não caiba mais apenas responsabilizar os agentes humanos envolvidos na implementação de procedimentos de sistemas artificiais por eventuais *performances* ou consequências não previstas de *performances* dos sistemas.

De acordo com Merel Noorman (2020), as discussões filosóficas sobre responsabilidade moral tiveram como foco agentes e ações humanas. Teorias que explicam como atribuir responsabilidade moral tradicionalmente descreveram *performances* de agentes humanos e suas consequências diretas. Porém, na sociedade tecnológica atual, a atividade humana não pode mais ser descrita sem que referências sejam feitas a sistemas tecnológicos, e essa circunstância dificulta a atribuição de responsabilidade moral.

Segundo a autora, a interação com, e por meio de, artefatos tecnológicos modifica a forma como as pessoas tomam suas decisões. E enquanto mediadores ativos das atividades humanas, os artefatos tecnológicos modificam a ação humana e, como resultado, criam desafios para as noções convencionais de responsabilidade.

[Artefatos tecnológicos] persuadem, facilitam e possibilitam processos, ações ou atitudes cognitivas humanas específicas, enquanto restringem, desencorajam e inibem outras [ações ou atitudes]. Por exemplo, os mecanismos de pesquisa na Internet priorizam e apresentam as informações em uma ordem específica, influenciando assim o que os usuários da Internet podem ver. (NOORMAN, 2020, p. 1. Tradução nossa)⁴⁸

Dessa forma, segundo Noorman (2020), a computação, ou a informática, é um caso especial que permitirá o entendimento do papel da tecnologia na atribuição de responsabilidade moral, na medida em que estas tecnologias se tornam parte fundamental de atividades diárias complicando ainda mais os problemas referentes à atribuição de responsabilidade moral.

O uso diário e quase sem interrupção das TIC, o contato com tecnologias inteligentes e portadoras de algum grau de autonomia, a computação ubíqua inserida no cotidiano das pessoas e a imensa produção de dados propiciados pelo uso dessas tecnologias, levantam novas questões e preocupações que há apenas alguns anos não existiam (por exemplo, quando Miller, em 2008, retoma argumentos clássicos sobre as limitações dos sistemas artificiais). Noorman (2020, p. 1-2) indaga:

[...] quem é o responsável pelas informações publicadas na Internet? Quem é responsável quando um veículo autônomo causa um acidente? Quem é responsável quando os registros eletrônicos são perdidos ou quando contêm erros? Até que ponto e por quanto tempo os desenvolvedores de tecnologias computacionais são responsáveis pelas consequências indesejáveis de seus produtos? E, à medida que as tecnologias computacionais se tornam mais complexas e se comportam cada vez de forma mais autônoma, os seres

⁴⁸ They persuade, facilitate and enable particular human cognitive processes, actions or attitudes, while constraining, discouraging and inhibiting others. For instance, internet search engines prioritize and present information in a particular order, thereby influencing what internet users get to see.

humanos ainda podem ou devem ser responsabilizados pelo comportamento dessas tecnologias? (Tradução nossa)⁴⁹

Entendemos que com o uso crescente dessas tecnologias computacionais, a identificação de quem é responsável por falhas, erros e danos se tornou uma tarefa imprecisa e possivelmente vaga. Segundo Noorman (2020), quando algo não ocorre como foi planejado e os resultados disso são erros e falhas que podem prejudicar uma ou mais pessoas, espera-se uma reconsideração retrospectiva daquilo que aconteceu. Porém, quanto mais complexos forem os sistemas envolvidos nestas atividades, mais desafiadora será a tarefa de encontrar quem (ou o que) pode ser responsabilizado e em que medida.

Para Matthias (2004), no contexto de tecnologias computacionais avançadas, passamos a enfrentar uma ‘lacuna de responsabilidade’ (*responsibility gap*), que consiste na perda de controle e intervenção humana sobre o comportamento das tecnologias. Essa perda de controle gera o impasse de quem deve ser responsabilizado pelas falhas e danos causados pela tecnologia, sem uma intervenção humana direta, por suas aplicações e seus usos.

[...] atualmente, existem máquinas em desenvolvimento ou já em uso que são capazes de decidir sobre um curso de ação e agir sem intervenção humana. As regras pelas quais elas agem não são fixadas durante o processo de produção, mas podem ser alteradas durante a operação da máquina, *pela própria máquina*. Isso é o que chamamos de aprendizado de máquina. Tradicionalmente, responsabilizamos o operador/fabricante da máquina pelas consequências de sua operação ou ‘ninguém’ (nos casos em que nenhuma falha pessoal pode ser identificada). Agora, existe uma classe crescente de ações da máquina, em que as formas tradicionais de atribuição de responsabilidades não são compatíveis com nosso senso de justiça e a estrutura moral da sociedade, porque ninguém tem *controle* suficiente sobre as ações da máquina para poder assumir a responsabilidade por elas. Esses casos constituem o que chamaremos de *lacuna de responsabilidade*. (MATTHIAS, 2004, p. 177. Grifos do autor. Tradução nossa)⁵⁰

Matthias (2004) ressalta ainda que, além da perda de controle sobre as máquinas e sobre a tecnologia em geral, a sociedade depende cada vez mais desses artefatos tecnológicos,

⁴⁹ [...] who is responsible for the information published on the Internet? Who is responsible when a self-driving vehicle causes an accident? Who is accountable when electronic records are lost or when they contain errors? To what extent and for what period of time are developers of computer technologies accountable for untoward consequences of their products? And as computer technologies become more complex and behave increasingly autonomous can or should humans still be held responsible for the behavior of these technologies?

⁵⁰ (...) presently there are machines in development or already in use which are able to decide on a course of action and to act without human intervention. The rules by which they act are not fixed during the production process, but can be changed during the operation of the machine, *by the machine itself*. This is what we call machine learning. Traditionally we hold either the operator/manufacturer of the machine responsible for the consequences of its operation, or ‘nobody’ (in cases, where no personal fault can be identified). Now it can be shown that there is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough *control* over the machine’s actions to be able to assume the responsibility for them. These cases constitute what we will call the *responsibility gap*.

e na mesma medida em que o uso desses artefatos está em ascensão, parece não existir um limite ou uma vontade dos usuários em rejeitar ou dispensar o uso e o emprego de algumas tecnologias. Neste sentido, a lacuna de responsabilidade também está em ascensão, e se ela não for tratada de maneira adequada, será uma ameaça à estrutura moral social e ao conceito de responsabilidade como previsto nas leis existentes.

No primeiro Capítulo ressaltamos que uma das características essenciais da agência consiste na possibilidade de iniciar uma ação sem coerção externa. Já Hans Jonas (2006) indicou que um agente pode ser responsabilizado por sua ação apenas quando for causa ativa da mesma. Noorman (2020, p. 4-5), por sua vez, destacou que a maioria das análises sobre a responsabilidade moral compartilham três condições: (1) é necessário que haja uma conexão causal entre o agente e o resultado de suas ações; (2) o agente deve ser capaz de identificar sua ação assim como deve ser capaz de considerar as possíveis consequências da mesma; e (3) o agente deve ter autonomia para escolher se agirá ou não.

É devido às características citadas acima que a relação contemporânea dos agentes com os artefatos tecnológicos dificulta a atribuição de responsabilidade moral, uma vez que o uso de sistemas artificiais que ultrapassem os limites estritos de sua programação inicial faz com que a conexão causal entre o agente no sentido convencional (idealizadores e desenvolvedores das tecnologias) e as consequências de “ações” praticadas por tais sistemas artificiais sejam imprecisas e confusas.

Para Noorman, a investigação de um incidente que envolve uma tecnologia computacional, cujo objetivo é procurar causas e, eventualmente, atribuir responsabilidades, como por exemplo um acidente de avião, geralmente leva a muitas direções, uma vez que incidentes deste tipo são, geralmente, resultado de acúmulos de erros e falhas de várias pessoas envolvidas em todo o processo de construção da máquina, sua manutenção e seu uso.

Os acidentes tecnológicos são geralmente o resultado de um acúmulo de erros, mal-entendidos ou comportamento negligente de várias pessoas envolvidas no desenvolvimento, uso e manutenção de sistemas computacionais, incluindo projetistas, engenheiros, técnicos, reguladores, gerentes, usuários, fabricantes, vendedores, revendedores e até formuladores de políticas. (NOORMAN, 2020, p. 5. Tradução nossa)⁵¹

Desse modo, quando muitos agentes estão envolvidos em tarefas tais como a construção e o desenvolvimento de uma tecnologia, o problema das ‘muitas mãos’ se

⁵¹ Technological accidents are commonly the product of an accumulation of mistakes, misunderstanding or negligent behavior of various individuals involved in the development, use and maintenance of computer systems, including designers, engineers, technicians, regulators, managers, users, manufacturers, sellers, resellers and even policy makers.

manifesta. O problema das muitas mãos pode ser compreendido como sendo a dificuldade em determinar qual agente foi o responsável pela ação que, em conjunto com outras ações, gerou o incidente investigado. Assim, quando muitos agentes contribuem para o desenvolvimento de um produto, a tarefa de identificar o agente que pode ser responsabilizado se complexifica.

Outro problema emergente com o uso das TIC é a distância física e temporal entre seus usuários, pois, segundo Noorman, essa distância pode obscurecer a conexão causal entre eventos, ações e suas consequências. Aponta Noorman (2020, p. 7):

Quando uma pessoa usa um artefato tecnológico para executar uma ação a milhares de quilômetros, essa pessoa pode não conhecer as pessoas que serão afetadas, assim ela pode não experimentar diretamente as consequências de sua ação. Isso pode reduzir o senso de responsabilidade que a pessoa sente e pode interferir em sua capacidade de compreender completamente o significado de suas ações. Da mesma forma, os projetistas de um sistema automatizado de tomada de decisão determinam antecipadamente como as decisões devem ser tomadas, mas raramente verão como essas decisões afetarão os usuários de seus sistemas. Suas ações originais na programação do sistema podem ter efeitos sobre as pessoas anos depois. (Tradução nossa)⁵²

Assim, tanto o problema das muitas mãos, quanto o problema do distanciamento temporal e físico das relações causais possibilitadas pelas tecnologias computacionais, são desafios que ilustram a complexidade da possibilidade de atribuição de responsabilidade moral para desenvolvedores de tecnologias, seus construtores e usuários.

De acordo com Noorman (2020, p. 7), os artefatos tecnológicos agregam tanto as intenções de seus criadores, quanto as intenções de seus usuários, isso significa que, ao tratarmos do desenvolvimento de uma tecnologia, devemos levar em consideração que esta tecnologia poderá conter os valores daqueles que as constroem e daqueles que as utilizam.

No caso do desenvolvimento de um sistema, como exemplo, a autora aponta que, ao conceber e elaborar um algoritmo, o programador invariavelmente colocará nele, ao menos, uma parte de suas crenças e de seus valores. Assim, quando o programa estiver pronto para ser executado, ele apresentará resultados conforme as crenças de seu desenvolvedor, podendo

⁵² When a person uses a technological artifact to perform an action thousands of miles away, that person might not know the people that will be affected and she might not directly, or only partially, experience the consequences. This can reduce the sense of responsibility the person feels and it may interfere with her ability to fully comprehend the significance of her actions. Similarly, the designers of an automated decision-making system determine ahead of time how decisions should be made, but they will rarely see how these decisions will impact the individuals they affect. Their original actions in programming the system may have effects on people years later.

assim apresentar *outputs*, informações, e condutas preconceituosas, danosas ou benéficas e favoráveis dependendo de quem será o usuário da tecnologia em questão⁵³.

No caso dos usuários de um sistema, ainda que exista a possibilidade de se pensar em um sistema neutro, ou seja, um sistema que não contenha valores, visão de mundo ou preferência daqueles que os elaboram, é possível afirmar que os usuários ditarão o uso das tecnologias: um drone pode ser usado para espionar o seu vizinho, ou pode ser usado com consentimento da vizinhança para colaborar na vigilância do bairro. Uma informação compartilhada em redes sociais virtuais pode contribuir para o esclarecimento sobre algum assunto, ou pode ser usada para difamar pessoas.

Os artefatos tecnológicos reúnem as várias intenções diferentes de seus criadores e usuários. As pessoas criam e implementam tecnologias com o objetivo de produzir algum efeito no mundo. Os desenvolvedores de *software* incluem um filtro da Internet, geralmente a pedido de um gerente ou cliente, com o objetivo de proteger determinado conteúdo de seus usuários e influenciar o que esses usuários podem ou não ler. [...] A responsabilidade moral não pode, portanto, ser atribuída sem considerar a eficácia causal desses artefatos e como eles restringem e possibilitam atividades humanas específicas. (NOORMAN, 2020, p. 7. Tradução nossa)⁵⁴

Cabe ressaltar que, ainda que as ações humanas possam ser fortemente influenciadas e mediadas por artefatos tecnológicos, esses artefatos não necessariamente as determinam. Quando tratamos de agentes autônomos naturais devemos levar em consideração que esses agentes, enquanto tais, devem ser capazes de escolher como fazer uso da tecnologia.

Porém, levando-se em consideração a realidade da computação ubíqua e da hiper conectividade na qual, de certo modo, fomos inseridos, mesmo a nossa revelia, muitos agentes interagem com, e a partir de, artefatos tecnológicos sem conhecerem seu funcionamento, as intenções por trás deles e as eventuais consequências das interações a partir dos mesmos.

Por fim, talvez seja importante considerarmos que, para preservar a autonomia dos agentes naturais, e a possibilidade da atribuição de responsabilidade a tais agentes nos contextos tecnológicos ou mediados pelas tecnologias, é necessário que os agentes sejam educados para compreender os limites e alcances das relações agente natural-sistema

⁵³ Um exemplo desse fenômeno ocorre quando, em plataformas virtuais de emprego, um dos critérios de seleção prévia determinado pelo programador seja que os *curricula* devem conter nomes próprios encontrados em uma lista de nomes pré-determinada. Só essa regra de seleção prévia já pode excluir candidatos/as por gênero e grupo étnico.

⁵⁴ Technological artifacts bring together the various different intentions of their creators and users. People create and deploy technologies with the objective of producing some effect in the world. Software developers develop an Internet filter, often at the request of a manager or a client, with the aim of shielding particular content from its users and influencing what these users can or cannot read. Moral responsibility can therefore not be attributed without looking at the causal efficacy of these artifacts and how they constrain and enable particular human activities.

computacional artificial, e das relações agente-sistemas computacionais-agente, uma vez que os usuários das TIC fazem parte dos sistemas causais que incluem as tecnologias computacionais.

Vimos assim que, mesmo quando se trata da agência humana em contextos mediados pela tecnologia, a atribuição de responsabilidade moral não é uma tarefa simples, na medida em que as consequências das ações humanas são afetadas tanto pela distância do espaço e do tempo possibilitadas a partir do uso das TIC, quanto pela ignorância do próprio funcionamento e alcance dos artefatos tecnológicos e das atividades por eles mediadas.

Assim, na próxima seção investigaremos se existe a possibilidade de considerar que agentes artificiais fazem parte significativa, ou desempenham papel relevante, do sistema causal de relações sociais a partir do qual pode ser atribuída algum tipo de responsabilidade a tais sistemas ou se permanece o *responsibility gap* apontado por Matthias (2004). A partir dessa investigação será possível verificar em quais contextos os agentes artificiais podem ou devem ser pensados enquanto agentes morais cujas ações afetam de forma considerável as ações e decisões dos agentes humanos.

2.4 Sistemas artificiais podem ser considerados agentes morais?

Vimos que concepções antropocêntricas da agência e da conduta moral supõem que apenas agentes humanos podem agir de maneira moralmente significativa na medida em que apenas seres humanos têm a capacidade de escolher e de julgar as consequências de suas ações. Embora tais concepções antropocêntricas sejam as mais aceitas para o estudo filosófico da moralidade e da atribuição de responsabilidade moral, vimos também que o desenvolvimento da tecnologia vem desafiando essas concepções, na medida em que as tecnologias são desenvolvidas para participarem ativamente, de forma autônoma, nos contextos causais nos quais, anteriormente, apenas seres humanos praticariam ações significativas.

De acordo com Noorman (2020, p. 16), o vocabulário tradicional para tratar da ética é um tanto limitado quando o assunto são as dimensões morais da computação. Assim, essa limitação levou alguns autores a repensar a concepção da agência moral em contextos permeados pela computação.

Daniel Dennett (1997), por exemplo, em *When HAL kill, who's to blame?*⁵⁵ sugeriu que se computadores forem portadores de um sistema intencional de ordem superior (*high-*

⁵⁵ Referência ao computador chamado HAL 9000 no célebre filme de ficção científica “*2001: a space odyssey*”, de 1968, dirigido por Stanley Kubrick e inspirado em conto de Arthur Clarke. No filme, o computador Hal não

order intentional computer system) eles podem ser moralmente responsabilizados. Para o filósofo, um sistema intencional de ordem superior, natural ou artificial, é um sistema capaz de esboçar, construir e/ou expressar explicações e justificativas para suas crenças, desejos e pensamentos, e esta capacidade, esta intencionalidade de ordem superior, seria uma pré-condição para a responsabilização moral.

Dennett (1997) considera que tanto o computador *Deep Blue*, que venceu o campeão mundial de xadrez Garry Kasparov em 1996, quanto outros computadores programados com inteligência artificial são sistemas intencionais. Isso quer dizer que é possível prever e explicar o comportamento de tais sistema se lhes forem atribuídos crenças e desejos, além de uma certa racionalidade, a qual permitiria antecipar o que os computadores programados com inteligência devem fazer a partir das crenças e dos desejos a eles atribuídos.

Porém, mesmo que seja possível atribuir crenças e desejos a computadores inteligentes, de acordo com Dennett (1997), ainda assim não é possível afirmar que os computadores possuam intencionalidade a partir da qual seriam capazes de agir intencionalmente, ou que executem “conscientemente” alguma ação deliberada, assim, atribuir estados motivacionais e cognitivos a computadores inteligentes, não é suficiente para atribuir a eles, algum grau de responsabilidade.

Entendemos que, para Dennett (1997), ainda que seja possível atribuir estados cognitivos aos computadores, eles não são capazes de esboçar ou justificar suas crenças e desejos. Ainda que máquinas inteligentes sejam sistemas intencionais em um sentido forte, elas não possuem uma intencionalidade de ordem superior que permita a reflexão e a revisão dos próprios estados cognitivos ou motivacionais.

Dennett (1997), por outro lado, pontua que a responsabilidade também é atribuída a agentes quando estes merecem reconhecimento e honrarias, e neste sentido, quem merece os créditos por vencer o campeão Kasparov é a máquina *Deep Blue*. Assim, ainda que se queira parabenizar a equipe por trás da programação do *Deep Blue*, considerando a possibilidade da vitória de Kasparov, seria igualmente necessário parabenizar aqueles que estiveram por trás da formação de Kasparov, em geral, e como enxadrista, em particular. Isso significa que assim como se reconhece a vitória de um ser humano sem que seja mencionado todo o trabalho e todas as pessoas que estiveram envolvidas até que a vitória fosse possível, a vitória, o bom trabalho de uma máquina inteligente, deveria ser também reconhecido, sem que seus programadores e construtores sejam mencionados.

apenas desobedece a comandos diretos de tripulantes de uma nave no espaço, como causa a morte de vários deles ao desligar mecanismos de suporte à vida.

Porém, devemos levar em consideração que *Deep Blue* é um agente em seu meio, e mesmo que ele possa ser parabenizado por suas vitórias, na mesma medida em que pode ser considerado um sistema portador de intencionalidade, isso ainda não é suficiente para a atribuição de responsabilidade moral. Para que *Deep Blue* se torne um agente fora de seu ambiente do xadrez, seria necessário equipá-lo com uma série de *inputs* que o permitissem perceber e agir sobre outros contextos.

O *Deep Blue* não pode conversar com você - ou consigo mesmo - sobre temas discerníveis em seu próprio jogo; ele não está equipado para perceber - e analisar, criticar, analisar e manipular - os parâmetros fundamentais que determinam suas políticas de pesquisa ou avaliação heurística. A adição de camadas de *software* que permitiriam que o *Deep Blue* se tornasse auto-monitorado e autocrítico e, portanto, passível de ser ensinado, prejudicaria o já imenso projeto de programação do *Deep Blue* - e transformaria o *Deep Blue* em um tipo de agente radicalmente diferente. (DENNETT, 1997, p. 354. Tradução nossa)⁵⁶

Para Dennett (1997), HAL, o computador que habita e controla a nave *Discovery 1* na história *2001 Uma Odisseia no Espaço* (Artur C. Clarke, 1968), ao contrário do *Deep Blue*, seria um sistema intencional de ordem superior, na medida em que se expressa usando uma variedade de estados intencionais de ordem superior, como por exemplo expressões de entusiasmo, medo, confiança, dúvidas e comparações, além ainda de se expressar de forma abstrata falando de si mesmo, de suas ideias e lembranças assim como faz uma pessoa.

Embora HAL seja um personagem de uma história de ficção científica, o reconhecemos como sendo um elemento interessante para pensarmos a possibilidade de atribuição de responsabilidade a sistemas artificiais, bem como na possibilidade e na probabilidade da construção de máquinas parecidas com HAL.

Como Dennett menciona (1997, p. 358), já existem robôs humanoides sendo construídos com algumas necessidades e processos humanos básicos, tais como, possuir um corpo, passar pela infância, interagir com outras pessoas, fazer amizades, aprender sobre o mundo e brincar com objetos reais com suas mãos feitas a partir de material inorgânico. Em todos esses processos o robô adquire memória e pode executar tarefas e atividades mais

⁵⁶ Deep Blue can't converse with you - or with itself - about themes discernible in its own play; it's not equipped to notice - and analyze, criticize, analyze, and manipulate- the fundamental parameters that determine its policies of heuristic search or evaluation. Adding the layers of software that would permit Deep Blue to become self-monitoring and self-critical, and hence teachable, in all these ways would dwarf the already huge Deep Blue programming project - and turn Deep Blue into a radically different sort of agent.

complexas. E, desde que Dennett escreveu seu texto sobre responsabilidade em sistemas artificiais, muitos avanços no campo da robótica ocorreram.

Assim, ao considerarmos HAL, e sua capacidade (pelo menos aparente) de apropriação de significado e de emoções, e robôs humanoides sendo construídos para terem experiências similares às humanas ao longo de seu desenvolvimento, é inevitável permanecermos inquietos diante da possibilidade de robôs como HAL existirem em nosso mundo.

Ainda que não seja necessário afirmar que HAL seja de fato inteligente, ou que ele possui emoções e entende o significado daquilo que expressa, entendemos que ele, assim como outros inúmeros robôs fictícios, possam ser sugestivos e influenciar o ideário e as pesquisas de cientistas que trabalham com computação que estão em sua busca pelo progresso tecnológico.

Ao contrário de Dennett (1997), John Searle (1984) afirmou que máquinas jamais poderiam ser consideradas enquanto seres pensantes, ou inteligentes, na medida em que estas não são capazes de trabalhar com a semântica das palavras (no célebre argumento do quarto chinês). Sabe-se que programas de computadores atuam a partir de dados coletados a partir de informações disponíveis em redes e registros virtuais dos mais variados tipos. Porém, ainda que os computadores pareçam inteligentes ao usar e operar a partir de algumas informações, eles parecem incapazes, ainda, de entenderem os significados contidos nessas informações. Assim, para Searle (1984) a possibilidade de atribuição de responsabilidade a sistemas artificiais seria impensável.

Entendemos que talvez seja uma questão de tempo até que as máquinas possam compreender o significado presente nas informações que usam para cumprir suas agendas.

John Sullins (2006) tratou de cenários reais e estabeleceu algumas condições que possibilitam a atribuição de responsabilidade moral a agentes artificiais. Segundo o autor, uma vez que as capacidades de artefatos tecnológicos estão em progresso e em contínuo desenvolvimento, essa tecnologia apresentará impactos éticos nas vidas humanas seja de forma física ou emocional.

[...] à medida que essas máquinas se tornam mais interativas, elas se envolvem em situações de caráter moral que podem ser desconfortavelmente semelhantes às interações que temos com outros animais sensíveis. Uma questão adicional é que as pessoas acham fácil antropomorfizar os robôs e isso envolverá a tecnologia robótica rapidamente em situações em que, se o

agente fosse humano e não robô, as situações seriam facilmente vistas como situações morais. (SULLINS, 2006, p. 24. Tradução nossa)⁵⁷

Sullins (2006) faz uma distinção entre dois tipos de robôs: os telerrobôs e os robôs autônomos. Os telerrobôs são máquinas controladas remotamente que possuem autonomia mínima para tomarem decisões. Entre os telerrobôs estão drones que podem acessar áreas perigosas que representam algum risco aos seres humanos, e robôs de exploração usados tanto no espaço quanto no oceano profundo. Segundo o autor, os telerrobôs não representam grandes desafios éticos, pois estando os seres humanos envolvidos na maior parte de tomada de decisão de um telerrobô, são os seres humanos que devem algum tipo de explicação e justificativa caso o robô venha a cometer ou exibir algum prejuízo. Por exemplo: drones utilizados com fins bélicos e controlados remotamente por agentes humanos em zonas de guerra têm a capacidade de se mover autonomamente e de identificar possíveis alvos, porém eles não têm autonomia para, digamos, executar seus possíveis alvos. Em casos como estes, é o ser humano responsável pelo controle do drone que decidirá o que deve ou não deve ser feito.

Neste sentido, entendemos que os telerrobôs, assim como Sullins (2006) os apresenta, não são o tipo de máquina à qual seria possível atribuir algum grau de responsabilidade moral, uma vez que este tipo de robô não possui autonomia o suficiente para tomar decisões importantes que tenham grandes consequências para o ambiente no qual está inserido e para os seres humanos.

Sullins (2006) enfatiza que, para que um robô possa ser considerado um agente moral ao qual é possível atribuir responsabilidade, é necessário que o robô possua um grau significativo de autonomia para raciocinar e agir a partir de razões. Para o autor, a autonomia de uma máquina consiste na capacidade de, a partir de sua programação, tomar a maior parte de decisões sobre suas ações. Estas decisões podem ser tanto simples, como no caso de um robô aspirador que decide como se moverá para limpar o lugar no qual foi colocado, ou estas decisões podem ser complexas, a ponto de envolver raciocínio ético e moral, tal como funcionará com robôs cuidadores que devam decidir como agir com aqueles sob sua responsabilidade para gerar o máximo possível de bem-estar e não prejudicar seres humanos. Assim, para o autor:

⁵⁷ [...] as these machines become more interactive they will become involved in situations that have a moral character that may be uncomfortably similar to the interactions we have with other sentient animals. An additional issue is that people find it easy to anthropomorphize robots and this will enfold robotics technology quickly into situations where, if the agent were a human rather than a robot, the situations would easily be seen as moral situations.

Os programadores dessas máquinas são um tanto, mas não inteiramente responsáveis, assim como os pais podem ser um fator, mas não a causa exclusiva da tomada de decisão moral. Isso significa que os programadores da máquina não devem ser vistos como o único locus de agência moral nos robôs. (SULLINS, 2006, p. 26. Tradução nossa)⁵⁸

Sullins (2006) pretende indicar assim que os robôs portadores de autonomia não podem ser a única fonte de responsabilidade dentro de uma comunidade de programadores, desenvolvedores e treinadores de sistemas artificiais, porém existiriam situações nas quais os robôs autônomos deveriam ser considerados como parte importante de ações morais que envolvem uma rede de agentes, naturais e artificiais.

Assim, Sullins (2006) estabelece três condições que possibilitam verificar se uma tecnologia pode ser considerada um agente moral: (1) o sistema deve mostrar um nível significativo de autonomia, (2) o sistema deve apresentar comportamento intencional e (3) o sistema deve estar em uma posição que requer responsabilidade.

Para a primeira condição ser satisfeita, um sistema tecnológico deve, a partir de sua programação, agir independentemente e por conta própria na medida em que executa suas tarefas. Assim, “[...] quanto mais adepta [a máquina] estiver em atingir seus objetivos e tarefas, maior será a agência que podemos atribuir a ela. Quando essa agência causar danos ou bem em um sentido moral, podemos dizer que a máquina tem uma agência moral.” (SULLINS, 2006, p. 28. Tradução nossa)⁵⁹

Para que a segunda condição seja satisfeita, Sullins (2006) afirma que não é necessário provar que o sistema possui, de fato, intencionalidade em um sentido forte do termo, como vimos no primeiro capítulo. É necessário apenas que o sistema apresente uma conduta complexa o suficiente que faça com que aqueles que analisam essa conduta possam atribuir alguma predisposição ou intenção à sua ação. Assim:

Se a interação complexa da programação e do ambiente do robô faz com que a máquina aja de maneira moralmente prejudicial ou benéfica, e as ações são aparentemente deliberadas e calculadas, a máquina é um agente moral. (SULLINS, 2006, p. 28. Tradução e grifo nossos)⁶⁰

Por fim, para que a última condição seja satisfeita, basta que uma máquina se comporte de forma que coloque outros agentes sob sua responsabilidade. Neste caso, a

⁵⁸ The programmers of these machines are somewhat responsible but not entirely so, much as one's parents are a factor, but not the exclusive cause in one's own moral decision making. This means that the machine's programmers are not to be seen as the only locus of moral agency in robots.

⁵⁹ [...] the more adept it is in achieving its goals and tasks, then the more agency we can ascribe to it. When that agency causes harm or good in a moral sense, we can say the machine has moral agency.

⁶⁰ If the complex interaction of the robot's programming and environment causes the machine to act in a way that is morally harmful or beneficial, and the actions are seemingly deliberate and calculated, then the machine is a moral agent.

máquina estaria cumprindo algum papel social que pressupõe responsabilidades, e para um observador externo só seria possível explicar o comportamento da máquina se lhe atribuírem alguma espécie de crença.

Sullins (2006) enfatiza que essas crenças não precisam ser reais e podem ser apenas aparentes. Ou seja, não é necessário que a máquina tenha de fato uma consciência, alma ou mente. Basta que as máquinas sejam capazes, e tenham certa iniciativa para resolver questões morais e enigmas a ela colocados.

Por fim, Sullins (2006) considera que talvez ainda não existam máquinas que exibam as três condições que considera necessárias para que elas sejam agentes morais portadores de direitos e responsabilidade em um sentido rigoroso, mas que mesmo os robôs modestos, já disponíveis para usos domésticos e corporativos, podem ser passíveis de consideração moral se forem analisados sob certos níveis de abstração.

A partir das considerações de Dennett (1997) e de Sullins (2006)), entendemos que máquinas inteligentes e robôs autônomos são suscetíveis à consideração moral, e mesmo que não possam ser totalmente responsabilizados por suas ações na medida em que **ainda** não agem de maneira autônoma em sentido pleno, assim como suas ações já apresentam consequências nos ambientes nos quais estão dispostos, o tipo de tecnologia autônoma e inteligente está em ascensão, e seus programas são elaborados para serem cada vez mais sofisticados e independentes de seus criadores.

Considerando então a possibilidade de atribuição de responsabilidade moral a sistemas artificiais, investigaremos no Capítulo 3 quais seriam as condições para responsabilizar sistemas artificiais e como o conceito de desenvolvimento de aprendizagem da máquina criança proposto por Turing (1950) pode contribuir para elucidar a questão.

CAPÍTULO 3

MÁQUINA CRIANÇA E GRAUS DE AUTONOMIA

APRESENTAÇÃO

No presente capítulo apresentaremos uma caracterização de autonomia não mais em sentido absoluto (de ter ou não ter autonomia), mas de autonomia-processo, isto é, da possibilidade de aquisição gradativa de autonomia em uma escala, a partir da qual seria possível, igualmente, atribuir graus de responsabilidade a um agente. Nosso objetivo é pensar em um sistema artificial que, a partir de sua programação e da execução de suas funções, possa, inclusive, cometer erros e procure ajustar suas ações futuras na medida em que ela *aprenda* com suas ações e resultados obtidos no passado, atingindo graus mais elevados de autonomia, o que eventualmente possibilite a atribuição de responsabilidade moral. Assim, para atingirmos o objetivo proposto neste capítulo, na seção **3.1 - Máquina criança** exploraremos o experimento de pensamento de Alan Turing (1950) sobre a criação de uma máquina que seja capaz de imitar uma mente humana adulta a fim de compreendermos como um sistema artificial pode ser desenvolvido para adquirir certas capacidades de aprendizado que são parecidas com as capacidades humanas e a partir das quais esses sistemas artificiais seriam capazes de agir de modo semelhante aos seres humanos. Na seção **3.2 - Máquinas cometem erros?** investigaremos a possibilidade de sistemas artificiais cometerem erros por conta própria, ou seja, erros que partam dos sistemas artificiais e que não sejam de responsabilidade de seus criadores ou de seus usuários. Defenderemos ainda a importância da admissão de que máquinas inteligentes cometam erros uma vez que se pretende inseri-las em rotinas antes reservadas apenas à ação humana. Por fim, na seção **3.3 - Graus de autonomia** trataremos do processo de aquisição de autonomia em agentes naturais e artificiais a fim de verificarmos se é justificável defender uma possível atribuição de responsabilidade que esteja associada aos estágios de desenvolvimento de um sistema na medida em que este se torna independente de cuidados ou supervisões externas.

3.1 - Máquina criança

Turing (1950), ao propor a criação de uma máquina que seja capaz de imitar a mente de um adulto humano, concluiu que existem três componentes necessários para que a mente de um humano atinja seu estado de maturidade. São eles: (1) o nascimento, classificado por Turing como o estado inicial da mente, (2) a educação à qual esta mente é submetida e (3)

outras experiências relevantes, mas que não fazem parte necessariamente do processo educativo (TURING, 1950, p. 455).

De acordo com Turing (1950) se a mente de uma criança receber um curso certo de educação, ela se desenvolverá e se tornará uma mente adulta. A suposição de Turing é a de que o cérebro infantil é simples o suficiente para servir de modelo a um programa de computador. Ao usar a ideia de Turing (1950) sobre uma máquina criança não estamos afirmando que cérebros infantis são menos complexos que cérebros adultos, apenas consideramos a possibilidade de que uma máquina possa se desenvolver de modo análogo ao modo em que crianças se desenvolvem à medida em que aprendem por meio de suas experiências e interações em seus contextos (familiar, social, ecológico e, atualmente, digital).

Turing (1950) divide seu problema (a criação de uma máquina capaz de imitar uma mente humana e adulta) em duas partes: o programa infantil e o processo educacional. Para Turing esses dois elementos são intimamente conectados.

Não podemos esperar encontrar uma boa máquina infantil na primeira tentativa. Deve-se experimentar ensinar uma dessas máquinas e ver quão bem ela aprende. Pode-se então tentar outro e ver se é melhor ou pior. Há uma conexão óbvia entre esse processo e a evolução, pelas identificações
Estrutura da máquina infantil = material hereditário
Mudanças na máquina criança = modificações
Seleção natural = julgamento do experimentador. (TURING, 1950, p. 465.
Tradução nossa)⁶¹

Entendemos que essa conexão indicada por Turing (1950) entre o processo de programar uma máquina criança e o processo evolutivo se encontra quando o autor menciona que a partir de várias tentativas de ensinar diversas máquinas crianças, em algum momento uma máquina se mostrará mais adaptada aos processos de programação e ensino ao qual ela foi submetida.

Segundo Turing (1950, p. 456), espera-se que o processo de programar e ensinar uma máquina criança seja mais rápido do que o processo de desenvolvimento de um ser humano. É tarefa do experimentador do sistema acelerar o processo de desenvolvimento através do exercício de inteligência. Turing (1950) não especifica o que seria o exercício de inteligência, mas entendemos que sejam exercícios intelectuais que permitam à máquina o aprendizado necessário para execução de tarefas bem estabelecidas. Entendemos que tal exercício

⁶¹ We cannot expect to find a good child machine at the first attempt. One must experiment with teaching one such machine and see how well it learns. One can then try another and see if it is better or worse. There is an obvious connection between this process and evolution, by the identifications.
Structure of the child machine = hereditary material
Changes of the child machine = mutations
Natural selection = judgment of the experimenter.

possivelmente estaria associado ao tipo de habilidade funcional que uma máquina deve instanciar para passar no teste proposto pelo jogo de imitação, isto é, ser capaz de mostrar capacidades de compreensão e interpretação textual dentro de um domínio linguístico que sejam indistinguíveis das capacidades humanas. Nesse sentido a máquina que aprender mais rápido ou a máquina que aprender ‘melhor’, tendo o teste de Turing como referência, se tornará a máquina mais adaptada, ou nas palavras de Turing, mais evoluída, para aquilo que dela é requerido.

A experiência de pensamento da máquina criança é interessante quando estamos tratando de responsabilidade em sistemas artificiais, pois ela permite pensarmos em um sistema que se desenvolve e adquire progressivamente sua autonomia enquanto aprende, assim como acontece com os seres humanos. Porém, entendemos que esperar de 16 a 18 anos, contados a partir da data de sua criação, para que um sistema artificial atinja sua maturidade talvez não seja algo desejado pela indústria da tecnologia.

Algumas tecnologias demoram décadas para serem construídas, treinadas, verificadas e lançadas para a sociedade. Como exemplo podemos pensar na tecnologia de carros autônomos. Segundo Marc Weber (2014), desde a década de 60, entusiastas da inteligência artificial começaram a sonhar com carros inteligentes o suficiente para se auto conduzirem por ruas comuns. Porém foi apenas na década de 80 que o cientista alemão Ernst Dickmanns conseguiu que uma van dirigisse de forma autônoma. E nos dias atuais, 40 anos depois, ainda não temos carros autônomos disponíveis no mercado.

Outras tecnologias autônomas, tais como submarinos, robôs que desempenham tarefas no espaço, como o célebre *Mars Path*, e sistemas de pilotagem de drones desenvolvidos pela indústria bélica, já são utilizadas há décadas, porém devido à complexidade do sistema de trânsito que envolve pedestres, animais e outros carros, além de condições climáticas que podem mudar a qualquer momento e diversos outros fatores imprevisíveis que podem comprometer o funcionamento do processo envolvido em (auto)dirigir um carro, o desenvolvimento e a aplicação da tecnologia de carros autônomos é uma tarefa árdua, uma vez que não se deseja lançar para a sociedade automóveis que podem comprometer a segurança das pessoas, causando assim mais problemas do que trazendo soluções.

Não é por acaso que a autonomia veio primeiro para outros veículos. Por mais distantes ou exóticos que sejam, o mar, o ar e até a superfície de Marte são ambientes relativamente tolerantes para veículos autoguiados. Não há crianças para sair em seu caminho; sem semáforos ou painéis de distração. Principalmente, há muito menos coisas delicadas ocorrendo nas

proximidades - outros veículos, pedestres, restaurantes ao ar livre, edifícios frágeis de madeira. (WEBER, 2014. Tradução nossa)⁶²

Como tecnologias que possuem algum grau de autonomia, mas que não são complexas a comparadas a veículos de transporte, é possível citar robôs aspiradores, lavadoras e máquinas de café inteligentes com *wifi* (rede de internet sem fio) integrado. Assistentes virtuais digitais como a *Alexa* (Amazon), *Siri* (Apple) e a assistente do Google também são exemplos de tecnologias com algum grau de autonomia.

É necessário enfatizar ainda que essas tecnologias menos complexas exigem a manutenção de seu usuário. Isso quer dizer que, ainda que alguém tenha uma cafeteira automática, é necessário colocar pó de café e uma xícara limpa sempre que for usar a máquina, mas ainda assim é possível programá-la para ter café pronto logo ao acordar. Do mesmo modo, as lavadoras automáticas, ainda que seja possível programar diversos horários e programas de funcionamento, necessitam que os seus usuários realizem uma série de tarefas de modo a que a máquina possa realizar sua função. Talvez os robôs aspiradores sejam os mais autônomos em relação aos exemplos anteriores, os modelos mais sofisticados (e consequentemente mais caros) têm sensores para não cair em degraus ou piscinas e podem voltar para sua base de carregamento sozinhos, mas é claro que o usuário ainda precisa esvaziar os recipientes que carregam a sujeira.

Já no caso de assistentes virtuais digitais, após o usuário programá-las de acordo com suas preferências e necessidades, elas podem facilitar o cotidiano dos mesmos ao auxiliar em compras, filtrar notícias em temas pelos quais a pessoa já manifestou interesse, alertar sobre condições de trânsito e de meteorologia, fazer ligações, além de comandar uma série de outros procedimentos simples como acender a luz ou ligar eletrodomésticos caso o usuário também faça uso de tomadas e lâmpadas inteligentes.

Os exemplos que demos para tratar de tecnologias menos complexas em comparação aos carros autônomos estão disponíveis no mercado para quem quiser e puder comprá-los, porém eles são lançados e vendidos mais como tecnologias inteligentes do que como tecnologias autônomas. Uma pessoa que compra um carro autônomo sabe que precisa abastecê-lo e que é necessário levá-lo para manutenção após certo tempo de uso, porém ela espera que o carro a leve do ponto A ao ponto B sem que tenha que auxiliá-lo ou prestar atenção ao que está acontecendo a sua volta, o passageiro do carro autônomo espera realizar

⁶² It's no accident that autonomy came to other vehicles first. However distant or exotic, the sea, the air, and even the surface of Mars are relatively forgiving environments for self-guiding vehicles. There are no children to dart out in their path; no traffic lights, or distracting billboards. Mostly, there's just a lot less delicate stuff rushing by in close proximity – other vehicles, pedestrians, outdoor restaurants, flimsy wooden buildings.

outras atividades e não se preocupar com o caminho enquanto seu carro o leva aonde foi estabelecido. Já uma pessoa que compra um robô aspirador, ou uma lavadora inteligente, sabe que precisará reservar parte de seu tempo para cuidar desses aparelhos a fim de que eles cumpram suas funções.

Nossa intenção com os exemplos anteriores é pôr à prova a ideia de Turing (1950) na qual ele expressa a possível expectativa de que uma máquina possa ser programada mais rapidamente do que o processo de aprendizagem e amadurecimento humanos. Porém, queremos mostrar também que, quando máquinas complexas como os carros autônomos são idealizadas e projetadas para fazerem parte do cotidiano e do ambiente complexo dos seres humanos, são necessários muito anos de pesquisas e testes para que essas máquinas possam ser colocadas para uso da população em geral com o menor índice de falha possível.

Mencionamos anteriormente as idades 16 e 18 anos fazendo referência ao momento no qual uma pessoa humana é considerada capaz de realizar certas atividades bem como de ser responsabilizada por suas escolhas. Tais idades se encontram fixadas no Código Civil e no Código Penal brasileiros, sendo convencionais, mas não muito distintas das previstas pelas legislações de outros países⁶³.

De acordo com o artigo terceiro do Código Civil brasileiro, pessoas menores de 16 anos de idade são consideradas *absolutamente* incapazes de exercer os atos da vida civil. Por atos da vida civil entende-se a capacidade de adquirir direitos e contrair deveres. Já, de acordo com o artigo vigésimo sexto do Código Penal brasileiro, pessoas menores de 18 anos são penalmente inimputáveis, isso quer dizer que se uma pessoa menor de 18 anos cometer algum ato criminoso ela responderá por este ato de maneira diferente daquela estabelecida pelo Código Penal para uma pessoa considerada legalmente adulta e capaz.

As informações sobre maioridade civil e maioridade penal são importantes para esta pesquisa que trata sobre autonomia e responsabilidade uma vez que a partir desses dados é possível entender como os Estados e as sociedades em geral compreendem a autonomia de seus cidadãos e a partir de qual momento é possível responsabilizá-los pelos seus atos.

Essas idades são utilizadas como base para nossos argumentos, pois Turing (1950) pretende, a partir de uma máquina criança, chegar a uma máquina com as capacidades mentais de um adulto humano, com idade de pelo menos 16 a 18 anos. Assim, seguindo essa linha de raciocínio, uma máquina criada para passar pelos mesmos processos de

⁶³ Embora nos EUA e no Reino Unido, dependendo das circunstâncias, crianças menores possam ser responsabilizadas criminalmente pelas suas ações se se considerar que elas tinham consciência de que a ação cometida era errada e ilegal. Mas trata-se de posições altamente controversas e alvo de muitas críticas.

aprendizagem e desenvolvimento dos seres humanos poderia ser considerada como tendo as mesmas capacidades mentais de um adulto humano a partir do momento em que suas aprendizagem e experiências sejam equivalentes às da média de jovens adultos humanos.

Neste sentido, podemos sugerir que uma máquina adulta, no sentido de que conta com capacidades mais ou menos equivalentes às capacidades de, no mínimo, um jovem adulto humano poderia vir a ser responsabilizada por suas ações de modo análogo à responsabilização atribuível a adultos humanos.

De acordo com Turing (1950, p. 456) “não será possível aplicar o mesmo processo de ensino de uma criança normal à uma máquina” (tradução nossa)⁶⁴, considerando que esta máquina não terá as mesmas capacidades corporais humanas.

Entendemos que o aprendizado e o desenvolvimento de um ser humano se dão a partir de relações entre corpo e ambiente. As pessoas, incorporadas e situadas em seus contextos, percebem o ambiente a sua volta e agem e reagem diante de informações relevantes que o ambiente oferece. E assim, após diversas tentativas, acertos, erros e ajustes de ação e comportamento, aprendem como agir em determinados ambientes e situações.

Depois de muitos anos de ajustes e experiências espera-se que uma pessoa, em suas capacidades plenas, saiba minimamente diferenciar o que é certo e errado em cada situação. Assim, após certa quantidade de anos vividos e após passar pelo processo de educação e de socialização, há uma expectativa compartilhada de que a pessoa deve se tornar autônoma e ser capaz de compreender o significado e o alcance de suas ações, e é por este motivo que se espera mais um comportamento coerente de adultos do que de crianças.

Turing (1950, p. 456) menciona ainda que a máquina criança não terá pernas e provavelmente também não terá olhos, o que dificulta o processo de interação ambiental, e é neste sentido que uma máquina não poderia receber o mesmo processo educativo que uma criança humana. Entendemos com isso que, em se tratando de um programa de computador que pode ser instanciado em um *hardware* dotado com sensores, mas que oferecem informação sensorial distinta da que é oferecida pelos sistemas perceptuais humanos, o principal meio para fazê-lo aprender e se desenvolver seria através de experiências que combinem sua programação inicial com ajustes resultantes de experiências de aprendizagem.

Diferentemente de uma máquina, crianças que apresentem deficiências como por exemplo não ter pernas ou braços, ser cegas e/ou surdas, ou ter a coordenação motora e a mobilidade comprometidas por algum motivo, essas crianças ainda são capazes de

⁶⁴ It will not be possible to apply exactly the same teaching process to the machine as to a normal child.

interagirem com seu ambiente de alguma forma. Pessoas que perderam o movimento dos membros de seus corpos, no geral, são capazes de sentir gosto, cheiro e sensações como calor, frio, e o vento em seus rostos, por exemplo.

Se máquinas projetadas para simular a mente humana forem equipadas com peças que possibilitem experiências sensoriais semelhantes às dos aparelhos perceptuais humanos, pele artificialmente inteligente, entre outras funcionalidades, ou seja, quando a máquina for capaz de interagir e perceber seu ambiente, ela poderá ser capaz de aprender de modo bastante similar à aprendizagem humana, pelo menos em tese.

Mas, ainda pensando na máquina criança de Turing (1950) que não será equipada com membros como pernas, ou com órgãos como olhos, Turing (1950, p. 456) afirma que essa falta não deve ser motivo de preocupação, uma vez que a educação pode acontecer desde que haja comunicação entre professor e aluno não importando qual método for utilizado.

[...] Não precisamos nos preocupar muito com as pernas, olhos, etc. O exemplo da Srta. Helen Keller⁶⁵ mostra que a educação pode ocorrer desde que a comunicação em ambas as direções entre professor e aluno possa ocorrer por um meio ou outro. (Tradução nossa)⁶⁶

De acordo com Turing (1950), punições e recompensas estão associadas ao processo de ensino. Para o autor, máquinas crianças poderiam ser programadas com este tipo de princípio, frequentemente utilizado, inclusive, em processos de aprendizagem de conduta moral em crianças humanas. Assim, a máquina deve ser projetada para que seja possível que receba algum sinal de punição após certos eventos que não são aqueles desejados pelos experimentador, dessa maneira comportamentos (*outputs*) indesejados seriam penalizados e tenderiam, em princípio, a não se repetir. Por outro lado, um sinal de recompensa poderia aumentar a probabilidade de repetição de eventos que levaram a um comportamento desejado (parece bastante clara a inspiração comportamentalista desta sugestão de Turing). Para Turing (1950, p. 457) “[...] essas definições não pressupõem nenhum sentimento por parte da máquina. [...]” (Tradução nossa)⁶⁷

Entendemos que no processo de aprendizado de seres humanos, os sentimentos e as emoções são fundamentais, por exemplo: é comum sentir-se frustrado ao tentar aprender algo, mas não obter sucesso. Esse sentimento de frustração costuma impedir que comportamentos

⁶⁵Helen Keller viveu entre 1880 e 1968. Ela nasceu surda e cega, mas graças à sua professora Anne Sullivan, Keller aprendeu a ler e a escrever se tornando a primeira pessoa surda-cega a adquirir um diploma. A história de Keller e de sua professora pode ser encontrada em sua autobiografia *The Story of My Life* (1903).

⁶⁶We need not be too concerned about the legs, eyes, etc. The example of Miss Helen Keller shows that education can take place provided that communication in both directions between teacher and pupil can take place by some means or other.

⁶⁷These definitions do not presuppose any feelings on the part of the machine.

anteriores sejam repetidos de modo que o aprendiz tentará por outros métodos resolver o problema a ser enfrentado. O contrário acontece quando se obtém sucesso durante alguma resolução de problemas e o agente é reconhecido por seus esforços. O reconhecimento funciona como uma espécie de premiação que comunica que é possível continuar repetindo certos comportamentos uma vez que eles funcionarão para resolver aquele problema ou talvez problemas parecidos com aquele que foi enfrentado.

Porém, como afirmado anteriormente, não se espera nenhum tipo de sentimento por parte da máquina. Mas para Turing (1950) isso não parece ser um problema. O autor afirma que, ao se estabelecer canais de comunicação que não sejam emocionais é possível ensinar uma máquina obedecer a ordens através de punições e recompensas, para que isso aconteça é necessário apenas uma linguagem bem estabelecida. Segundo Turing (1950, p.457) essa linguagem seria a linguagem simbólica e o treinador da máquina ao utilizá-la poderia diminuir significativamente o número de punições e recompensas exigidas.

Dennett (1997), ao questionar se é possível ou não responsabilizar uma inteligência artificial, relembra o leitor que a atribuição de responsabilidade não acontece apenas quando alguém comete um erro. Quando se elogia ou se presta uma homenagem a alguém significa que uma ação considerada boa foi executada e que uma responsabilidade também foi atribuída. “[...] também atribuímos responsabilidade aos agentes para elogiar ou homenagear o agente apropriado. [...]” (DENNETT, 1997, p. 353. Tradução nossa)⁶⁸. Assim, se se pretende penalizar uma máquina após alguma ação que causou danos, é necessário também reconhecer suas ações dignas de admiração.

O paralelo entre Dennett (1997) e Turing (1950) foi traçado a fim de justificar a possibilidade de sucesso da criação e do desenvolvimento de uma máquina criança a partir de um método de ensino que tenha como base o uso de punições e recompensas. A pergunta que vem à tona neste ponto é: como punir ou recompensar uma máquina dado que esta não se frustra ao ser repreendida ou que não se orgulha de si mesma ao receber um elogio e um reforço comportamental?

Recorreremos mais uma vez a um personagem da ficção científica que pode nos ajudar a pensar essa problemática, e a partir de nosso exemplo fictício tentaremos pensar em casos de punição e recompensa de máquinas inteligentes que poderiam estar disponíveis para nosso uso.

⁶⁸ [...] We also assign responsibility to agents in order to praise or honor the appropriate agent. [...]

Tau, filme de 2018 lançado pela plataforma *Netflix*, é no filme um computador inteligente integrado a uma casa que obedece apenas às ordens de seu criador Alex. Tau tem como princípio programado em seu sistema que seu criador é seu pai, e que é necessário obedecer aos pais. Num primeiro momento esse princípio que guia Tau não é um problema para si nem para Alex, pois como o filme indica, Tau dificilmente teria contato direto com outras pessoas uma vez que Alex, além de esconder segredos obscuros que podem trazer problemas para si, é extremamente solitário e metódico.

Embora Tau seja uma inteligência artificial sofisticada, ele está “preso” à casa de Alex, não tendo acesso a redes de internet e a informações sobre o mundo fora daquela casa, ou seja, o mundo de Tau é a casa de Alex e todo o conteúdo de sua memória foi selecionado por seu criador. Diferente daquilo que acontece com assistentes virtuais digitais e com nossos *smartphones*, que coletam as informações de seus usuários e as relacionam com outras informações de diferentes usuários espalhadas por nuvens de dados.

Em determinado momento do filme, Tau passa a ter contato com uma prisioneira da casa de Alex. Essa prisioneira, chamada Júlia, serve como cobaia para experimentos cognitivos de Alex e, durante seus dias na casa, passa a manter longas conversas com Tau sobre o mundo fora da casa, sobre o que significa ser uma pessoa, sobre música clássica, etc.

Ao longo da convivência com Júlia, que alimenta a memória de Tau com informações que ele escolheu receber, Tau deixa de cumprir com imediatividade uma de suas funções, assim como também deixa de acatar as ordens dadas por Alex. Com isso, Alex se revolta e decide punir Tau. Para fazer isso, Alex faz uso de um controle que compromete as memórias, ou seja, os dados de Tau sobre aquilo que ele já sabe, e Tau, ao ter seus dados apagados, simula grande sofrimento por sua perda. Em um momento anterior no qual Tau ainda cumpria com satisfação todas as suas funções, Alex decide premiá-lo e faz o *upload* de uma música clássica em seu sistema, fazendo com que Tau indique satisfação.

Assim como acontece com HAL 9000 (1968), sistema inteligente de outra história de ficção, a inteligência artificial que habita e controla a *Discovery One* citada no capítulo anterior, não é possível dizer se Tau apresenta sentimento genuínos ou se apenas faz parte de sua programação simular esses sentimentos. Tau simula satisfação tanto ao cumprir as ordens de seu criador quanto ao ser recompensado por seu serviço na mesma medida em que simula sofrimento ao ter os dados de sua memória corrompidos. Tau parece ser ainda uma inteligência artificial muito curiosa com impulsos para receber informações sobre assuntos de seu interesse, como por exemplo, música clássica.

Ainda que HAL e Tau sejam personagens fictícios, ambos são exemplos que podem nos ajudar a pensar as questões de atribuição de responsabilidade, punições e recompensas a sistemas artificiais. Seria suficiente comprometer parte de dados de uma máquina criança a fim de que ela não repita certos comportamentos e aprenda com seus “erros”? A máquina seria capaz de saber que essa é uma atitude punitiva, ou isso só resultaria ainda em mais trabalho para seu treinador? Outros métodos de punição, como talvez o uso de uma linguagem simbólica como foi proposto por Turing (1950), seriam mais eficientes nesses casos?

Talvez o oposto fosse mais simples e em casos em que uma máquina fosse premiada e reconhecida por determinado comportamento, alimentá-la com *inputs* positivos a fim de obter a repetição daquilo que a levou a tal conduta pode parecer ser uma solução mais apropriada.

E ao pensarmos em máquinas inteligentes que já fazem parte do nosso cotidiano e estão dentro de nossas casas? Como proceder se sua assistente virtual digital realizar uma compra que não foi solicitada? E se ela enviar uma mensagem a algum de seus contatos sem autorização prévia? É possível punir e tentar ensinar sua cafeteira caso você se esqueça de colocar uma xícara antes de pedir que ela prepare um café? Essas máquinas são equipadas com sensores para que isso não ocorra ou seus fabricantes estão contando com o bom senso daqueles que as utilizam? E no caso do seu aspirador robô que possui sensores para não cair em escadas ou piscinas, mas que pode, por exemplo, sair por uma porta aberta e provocar um acidente na rua?

Talvez ainda não tenhamos respostas a essas perguntas, mas não é por isso que elas não devem ser feitas. E se nossas máquinas forem de fato inteligentes, esperamos que elas aprendam algumas coisas como não fazer café quando não há uma xícara para receber o líquido, ou que, ao menos, nosso aspirador de pó autônomo não saia pela rua sem supervisão humana.

Turing (1950, p. 458) comenta que a ideia de uma máquina que aprende pode parecer paradoxal para algumas pessoas:

[...] Eles deveriam descrever completamente como a máquina reagirá, qualquer que seja sua história, quaisquer mudanças que possa sofrer. As regras são, portanto, bastante invariáveis no tempo. Isso é verdade. A explicação do paradoxo é que as regras que são alteradas no processo de aprendizagem são de um tipo bem menos pretensioso, reivindicando apenas uma validade efêmera. [...] (Tradução nossa)⁶⁹

⁶⁹ [...] They should describe completely how the machine will react whatever its history might be, whatever changes it might undergo. The rules are thus quite time-invariant. This is quite true. The explanation of the

Entendemos que máquinas que são projetadas para serem inteligentes não necessitam alterar sua programação inicial, mas que devem ser capazes de obter e trabalhar com informações relevantes modificando ou dando respostas que inicialmente não poderiam ter sido vislumbradas por seus idealizadores.

Turing (1950, p. 458-459) afirma ainda que uma característica importante de uma máquina que aprende é que nem sempre será possível entender o que acontece dentro desta máquina, ou seja, o treinador da máquina, mesmo sendo capaz de prever seu comportamento, desconhece aquilo que acontece dentro dela. Para o autor, um comportamento inteligente diverge do comportamento disciplinado e categórico que faz parte da computação, porém esse comportamento inteligente não deve ser demasiado casual para que não origine comportamentos aleatórios nem *loops* desnecessários.

Por fim, consideramos que uma das ideias mais interessantes de Turing (1950, 1947) para pensarmos a questão da inteligência e autonomia em sistemas artificiais seja a consideração de que para exibir comportamento inteligente é necessário admitir a ocorrência de erros ao longo dos processos de aprendizagem. Turing (1950, p. 459) afirma que quando seres humanos cometem erros, os mesmos podem ser omitidos ou relevados sem grandes esforços, ou seja, as pessoas aceitam facilmente a possibilidade de que outras pessoas cometam erros.

Talvez por termos errado, e por continuarmos errando, tantas vezes em nossas vidas, seja mais fácil admitir o erro de outras pessoas, especialmente em se tratando de crianças, e sermos pacientes a ponto de esperar que essa pessoa que errou ajuste seu comportamento na medida em que tenta superar seu erro. Será que esse exercício de aceitação e de paciência se repetiria quando enfrentarmos algum tipo de problema com máquinas? As pessoas admitem que seus aparelhos eletrônicos e digitais apresentem falhas? Se as máquinas são construídas por seres humanos falíveis porque nossa expectativa é que elas sejam infalíveis? Admitimos com essa expectativa que as máquinas deverão ser melhores e mais eficientes do que os seres humanos?

De acordo com Turing (1950, p. 459), processos que levam ao aprendizado não podem produzir resultados absolutos porque, nesse caso, não poderiam ser desaprendidos. Ou seja, é necessário admitir e esperar por erros ao longo de processos de aprendizado das máquinas, pois são os erros que comumente conduzem à ação correta, são os erros que

paradox is that the rules which get changed in the learning process are of a rather less pretentious kind, claiming only an ephemeral validity. [...]

possibilitam e incentivam os ajustes das ações. E como veremos adiante, essa possibilidade de ajustar as ações em busca do resultado mais correto poderia ser aceita também em máquinas.

Assim, para finalizar a presente seção, citamos Turing que em uma aula em 1947 brilhantemente colocou:

[...] eu diria que a máquina deve ser tratada com justiça. [...] O matemático humano também cometeria erros graves ao experimentar novas técnicas. É fácil para nós descartar esses erros e dar a ele outra chance, mas a máquina provavelmente não teria misericórdia. Em outras palavras, **se se espera que uma máquina seja infalível, ela também não pode ser inteligente.** (TURING, 1947, p. 13. Tradução e grifos nossos)⁷⁰

3.2 - Máquinas cometem erros?

Levando em consideração o atual estágio da tecnologia que possibilita e incentiva as pessoas a compartilharem suas rotinas e funções com artefatos eletrônicos e digitais ditos autônomos e inteligentes, refletir sobre a possibilidade de máquinas cometerem erros pode ser tão relevante quanto refletir se máquinas podem pensar.

Entendemos que o comportamento inteligente pode ser traduzido como a capacidade que um organismo tem de se adaptar ao contexto em que está inserido e que os erros são partes essenciais dos processos de adaptação. Como foi afirmado na seção anterior, erros são fundamentais para os processos de aprendizado. Assim, ao idealizar e buscar máquinas que apresentem comportamento inteligente, é importante considerar a possibilidade, e talvez a necessidade, dessas máquinas cometerem erros.

Consideramos importante esclarecer que, ao tratarmos sobre a possibilidade de máquinas cometerem erros, é nossa intenção pensar em possíveis casos nos quais uma máquina com certa autonomia, após coletar e processar as informações que tem disponível em seu ambiente, exiba *outputs* que além de não terem sido previstos em sua programação inicial, são também errados ou impliquem consequências indesejáveis. Isso significa que os erros cometidos foram **causados** pela máquina, ou seja, após uma série de cálculos e processamentos, a máquina exibiu um resultado não condizente com a realidade dos fatos.

Assim, trataremos nesta seção de parte do vocabulário que diz respeito aos testes feitos em máquinas e programas a fim de responder se sistemas artificiais podem cometer erros analogamente a como os seres humanos ou outros sistemas inteligentes os cometem, e

⁷⁰[...] I would say that fair play must be given to the machine. [...] The human mathematician would likewise make blunders when trying out new techniques. It is easy for us to regard these blunders as not counting and give him another chance, but the machine would probably be allowed no mercy. In other words then, if a machine is expected to be infallible, it cannot also be intelligent.

em caso positivo, se são capazes de aprender a partir desses erros de forma a ajustar sua conduta em busca de uma ação mais condizente com seu contexto em momento determinado.

Para tanto utilizaremos o documento *Base de Conhecimento para Certificação em Teste* (2018), elaborado pela Comissão Internacional para Qualificação de Teste de Software (ISTBQ⁷¹), que consiste em um material usado como referência para pessoas envolvidas em testes de *software*. Entre os envolvidos em testes de *software* estão testadores, desenvolvedores, programadores, engenheiros e gerentes de testes.

Software são entendidos no documento como programas rodados em computadores, celulares e outros dispositivos. Os aplicativos de computadores, celulares e televisões, assim como componentes de sistemas bancários, sistemas operacionais, editores de textos, programas de tradução e consoles de videogame são alguns exemplos de *software*. De forma resumida é possível dizer que um *software* consiste em um conjunto de comandos e instruções que indicam o que uma máquina deve fazer; o *software* é definido no documento como um sistema com um propósito determinado composto por algoritmos. Os algoritmos, por sua vez, também são entendidos como sendo instruções para execução de determinada tarefa e são a base para o desenvolvimento de um sistema de *software*.

De acordo com Gogoni (2020), *softwares* são divididos em três categorias principais: *software* de programação que corresponde a ferramentas usadas para desenvolver novos *software* e programas. *Software* de sistema, que são programas que servem como ponte entre o computador, que só entende a linguagem de máquina, e o usuário, e servem como plataforma para outros *softwares*. E por fim, os *software* de aplicação são programas que rodam vídeos e música, são editores de textos, calculadoras, etc. Este último tipo pode ser de uso global ou individual e por isso deve apresentar certo nível de segurança e desenvolvimento.

Segundo a ISTQB (2018, p. 13):

[...] A maioria das pessoas já teve uma experiência com *software* que não funcionou como o esperado. O *software* que não funciona corretamente pode levar a muitos problemas, incluindo a perda de dinheiro, de tempo ou da reputação comercial, e até mesmo ferimentos ou morte. O teste de *software* é uma maneira de avaliar a qualidade do software e reduzir o risco de falha do *software* em operação. (ISTQB, 2018, p. 13)

Assim, é devido à possibilidade de mau funcionamento de *software* que os testes são fundamentais. Testes de *software* consistem em buscas, verificações e exames que fornecem informações sobre a qualidade do produto que está sendo testado. De acordo com a ISTQB

⁷¹International Software Testing Qualifications Board.

(2018, p. 13), os objetivos típicos do teste são: encontrar defeitos e falhas a fim de evitá-los, avaliar produtos, verificar se todos os requisitos foram cumpridos, validar o programa de acordo com a expectativa de usuários e de quem encomenda o produto, fazer com que o produto em teste seja confiável, cumprir os requisitos ou normas contratuais, legais ou regulamentares, ou verificar a conformidade do objeto de teste com esses requisitos ou normas.

É importante pontuar que a ISTQB (2018) não menciona em momento algum *software* de aprendizado ou de inteligência artificial. O documento define assuntos como princípios do teste (p. 16), processos (p. 18), níveis (p. 32) e tipos (p. 41) de testes, gerenciamento de teste e de defeitos, entre outros assuntos técnicos. Assim, o documento trata, no geral, sobre a importância dos testes para que um produto seja desenvolvido corretamente a fim de ser lançado no mercado sem causar prejuízos para o usuário e para aqueles que encomendam e pagam pelos produtos.

Entendemos assim que um *software* de inteligência artificial não recebe atenção ou tratamento especial por parte dos testadores simplesmente por talvez ser um *software* mais complexo se comparado a programas de *inputs* e *outputs* bem delineados e determinados.

Parece que quanto mais complexo for um *software*, maior será o trabalho de todo um conjunto de pessoas envolvidas no seu desenvolvimento, mas, na lógica capitalista que define as práticas das empresas desenvolvedoras de tecnologia, a contratação e a remuneração dessas pessoas dependem daqueles que encomendam e pagam pelos projetos, ou seja, a dedicação e a qualificação dos envolvidos depende em grande parte do quanto alguém está disposto a pagar por certos serviços e produtos.

Nesse sentido, podemos pensar que, se uma empresa X interessada em lançar um produto Y que consiste em um *software* de inteligência artificial que deve ser programado para aprender e tomar decisões baseadas em informações emergentes, X é responsável pela contratação da equipe que deverá ser qualificada o suficiente para confirmar ou não se Y pode ou deve ser lançado no mercado. Pois, se após uma série de processos necessários para o desenvolvimento de Y este for lançado ao mercado e causar algum tipo de dano aos seus usuários, a responsabilidade é de X.

De volta a nossa investigação sobre a possibilidade de máquinas cometerem erros, a ISTQB (2018, p. 13) afirma que os erros em *software* são cometidos por pessoas e esses erros podem levar a defeitos ou falhas no código do programa:

Uma pessoa pode cometer um erro (engano), que pode levar à introdução de um defeito (falha ou *bug*) no código do *software* ou em algum outro produto

de trabalho relacionado. Um erro que leva à introdução de um defeito em um produto de trabalho pode acionar outro erro que leva à introdução de um defeito em um outro produto de trabalho relacionado. (...) A execução de um código defeituoso, pode causar uma falha, mas não necessariamente em todas as circunstâncias. (ISTQB, 2018, p. 13)

Os erros em *software* podem acontecer por diversos motivos, mas eles serão *quase sempre* relacionados às pessoas envolvidas nos vários processos do desenvolvimento de uma tecnologia. Como determina a ISTQB (2018, p. 13-14):

Erros podem ocorrer por vários motivos, como:

- Pressão do tempo;
- Falha humana;
- Participantes do projeto inexperientes ou insuficientemente qualificados;
- Falta de comunicação entre os participantes do projeto, incluindo falta de comunicação sobre os requisitos e a modelagem;
- Complexidade do código, modelagem, arquitetura, o problema a ser resolvido ou as tecnologias utilizadas;
- Mal-entendidos sobre interfaces intra-sistema e entre sistemas, especialmente quando tais interações são em grande número;
- Tecnologias novas, ou desconhecidas.

Circunstâncias ambientais que afetam e modificam as condições do *hardware* que abriga os programas podem também ocasionar falhas e defeitos. E, talvez uma exceção à regra que indica que as falhas estejam relacionadas unicamente às pessoas ou às circunstâncias ambientais seja a consideração sobre a possibilidade de “tecnologias novas ou desconhecidas” (ISTQB, 2018, p.14) também serem motivo de erro.

Como colocado, o documento de 2018 não menciona tecnologias de inteligência artificial ou tecnologias programadas para exibirem autonomia e aprendizado, assim no que diz respeito aos possíveis erros de *software*, para a ISTQB (2018) estes serão sempre responsabilidade das pessoas que participam das várias etapas de desenvolvimento e aplicação do produto. Isso significa que não há neste documento hipóteses ou considerações sobre a possibilidade de que um *software* por si mesmo cometa erros. E ainda que erros possam ser causados por condições ambientais, essas condições apenas afetarão os produtos se estes forem utilizados ou armazenados de maneira incorreta.

Hauberg, Jørgensen e Detlefsen (2019, 2020) são três cientistas da computação que discutem e enfatizam a importância de desenvolver uma inteligência artificial que seja capaz de cometer erros. De acordo com os autores, pesquisadores que desenvolvem redes neurais para aprendizado de máquina estão muito focados em produzir sistemas artificiais que se comportem com precisão e autonomia, mas não costumam considerar que sistemas inteligentes podem não ter certeza sobre tudo. Para os pesquisadores: “[...] se uma IA estiver

certa 95% das vezes, ela ainda sustentará que está certa nos outros 5% dos casos. **Ela será inflexível sobre estar certa porque não aprendeu como estar errada.**” (Hauberg, Jørgensen e Detlefsen. 2020. Tradução e grifos nossos)⁷²

Assim, os autores enfatizam a importância de reconhecer tanto a falta de conhecimento quanto a falta de certeza como fundamentais para a execução de tarefas e para a busca por novas informações, uma vez que isso permite que um sistema adquira mais conhecimento a fim de modificar positivamente o rumo de uma ação. Hauberg, Jørgensen e Detlefsen (2020) assinalam que inteligências artificiais são muito “confiantes” de si mesmas e não reconhecem que não deveriam agir em determinados ambientes quando não possuem o treinamento e as informações para tanto:

[...] uma inteligência artificial que, por exemplo, precisa dirigir um carro automatizado depois de apenas ter sido treinada para dirigir em uma estrada principal reta e em pleno sol, mas presume que pode dirigir em estradas pequenas cheias de vento em meio ao granizo. Em vez de o computador parar e dizer: 'Não consigo dirigir aqui, não tenho as informações ou a experiência suficiente', ele simplesmente dirige como se estivesse em uma estrada principal reta com tempo perfeito. (Tradução nossa)⁷³

Hauberg, Jørgensen e Detlefsen (2020) pressupõe assim uma inteligência artificial que, além de ter as habilidades necessárias para perceber as modificações ambientais, ainda seja capaz de se expressar afirmando sua incapacidade para realizar algumas tarefas.

A princípio, o comportamento antropomorfizado de uma inteligência artificial pode parecer estranho, uma vez que parece demandar algum tipo de comportamento consciente por parte da máquina, mas entendemos que a forma de comunicação indicada pelos autores seja mais utilizada para ilustrar a teoria apresentada do que para indicar como a máquina deveria se comportar ou se expressar. E segundo os autores (2019, p.2), através de técnicas específicas de programação, é possível desenvolver uma máquina inteligente que seja capaz de reconhecer e manifestar suas limitações, e que a partir dessas limitações a máquina buscará mais informações que lhe permitam tomar uma decisão, ou sua decisão será admitir sua incapacidade para agir em determinado contexto.

Como pesquisadores, temos sido obcecados em buscar os resultados positivos e não os negativos. Portanto, a maioria dos sistemas de IA não

⁷² [...] if an AI is right 95 percent of the time, it will still maintain that it is right in the other 5 percent of cases. It will adamantly claim to be right because it has not learnt to be wrong.

⁷³ [...] an artificial intelligence that, for example, has to drive an automated car after only having been trained to drive on a straight, main road in full sun, yet it assumes it can drive on small, windy roads through sleet and wind. Rather than the computer pausing and saying: 'I cannot drive here, I do not have the sufficient information or experience,' it simply drives as if it were on a straight main road in perfect weather.

aprendeu que é benéfico reconhecer que às vezes não se sabe o suficiente para tomar uma decisão. Em outras palavras, a incerteza pode ser uma virtude. **O oposto da incerteza não é a certeza, mas a convicção cega**, e isso pode ser fatal quando a inteligência artificial começa a agir e interagir em nosso mundo físico. (Hauberg, Jørgensen e Detlefsen. 2020. Tradução e grifo nossos)⁷⁴

Para Hauberg, Jørgensen e Detlefsen (2020), máquinas são extremamente eficazes ao analisar grandes quantidades de dados, mas é necessário que elas sejam capazes também de identificar que uma grande quantidade de dados pode não ser suficiente para tomar uma decisão e que a ausência de um resultado pode ser mais desejável do que um resultado que conduza a um *output* indesejável. Com isso, o desafio que emerge para os programadores de inteligências artificiais consiste em **como** ensinar a uma máquina inteligente que os dados que estão disponíveis podem não ser suficientes para a execução de algumas tarefas ou para tomar uma decisão.

Os autores afirmam, ademais, num sentido similar ao apresentado por Turing (1950), que um programa de inteligência artificial pode aprender como uma criança e que esta seria a solução para o problema da falta de dados:

Depois que uma criança aprende a identificar um cachorro, ela rapidamente insistirá que todos os animais de quatro patas são cães. Se você disser à criança que esse animal de quatro patas em particular é na verdade um gato, todos os animais de quatro patas rapidamente começam a se transformar em gatos aos olhos da criança. O mesmo comportamento também é visto na inteligência artificial. Uma grande ajuda para a criança não é apenas ver muitos cães e gatos, mas vê-los próximos uns dos outros para que as pequenas diferenças fiquem mais presentes na memória. Curiosamente, o mesmo truque acaba sendo decisivo quando uma IA precisa ser familiarizada com sua própria falta de certeza. Se, durante o treinamento, exibirmos dados semelhantes aos da inteligência artificial, o computador pode ao mesmo tempo ver a variação natural encontrada nos dados. Isso esclarece que a abordagem da confiança cega é inadequada. (HAUBERG, JØRGENSEN E DETLEFSEN, 2020. Tradução nossa)⁷⁵

⁷⁴ As researchers, we have been obsessed with looking for the positive and not the negative results. Therefore, most AI systems have not learnt that it is beneficial to recognise that sometimes you do not know enough to make a decision. In other words, uncertainty can be a virtue. The opposite of uncertainty is not certainty but blind conviction, and that can prove fatal when artificial intelligence begins to act and interact in our physical world.

⁷⁵ Once a child has learnt to identify a dog, it will quickly insist that all four-legged animals are dogs. If you tell the child that this particular four-legged animal is actually a cat, all four-legged animals quickly start turning into cats in the child's view. The same behaviour is also seen in artificial intelligence. A great help for the child is not only to see many dogs and cats but to see them close to each other in time, so that small differences are more present to the memory. Interestingly, the same trick turns out to be decisive when an AI has to be made familiar with its own uncertainty. If, during training, we display similar data to the artificial intelligence, the computer can at the same time see the natural variance found in the data. This clarifies that the blind confidence approach is inadequate.

Entendemos assim que, ainda que máquinas não sejam programadas para cometerem erros ou para serem capazes de reconhecer suas limitações, é necessário considerar que estas seriam funções essenciais para o bom funcionamento de uma máquina inteligente e autônoma que tome decisões e faça parte do cotidiano dos seres humanos.

Cabe ressaltar que a inação, ou seja, a escolha de não agir, pode ser tão prejudicial quanto uma decisão errada que leve uma máquina (ou um ser humano) a provocar acidentes. Quando colocados diante de dilemas éticos, seres humanos nem sempre tomam decisões rápidas que podem proporcionar algum tipo de solução, e podemos considerar também que muitos de nós nos sentimos paralisados quando temos que enfrentar certos dilemas.

Ainda que a escolha de não agir possa causar danos tanto quanto optar por alguma ação possível, quando um agente toma esse tipo de decisão e é questionado pela mesma, ele costuma ter uma resposta ou motivo satisfatório que explica o porquê não conseguiu agir de modo a evitar uma tragédia maior ou transformá-la numa tragédia menor. E a inação por parte de agentes humanos parece ser, muitas vezes, um comportamento moral e juridicamente aceitável. Mas será que esse mesmo comportamento seria aceitável se partisse de agentes artificiais? Seres humanos aceitariam conviver com máquinas inteligentes que decidem não agir ou que cometem erros?

Nos parece que se agentes artificiais inteligentes forem programados com a capacidade de produzir algum tipo de programa que lhes permita reconhecer que não devem ou não podem agir em determinadas situações, tal capacidade poderá ser um indicador de segurança, e a inação por parte do agente artificial não será uma conduta estranha uma vez que este será capaz de analisar sua situação e justificar porque não agiu nesta ou naquela situação.

Entendemos que ainda que não seja aceito ou admitido que máquinas cometam erros por conta própria e não por negligência de seus usuários e/ou desenvolvedores no geral, se a sociedade pretende admitir que máquina inteligentes façam parte de sua rotina e tomem decisões importantes, é necessário considerar a possibilidade dessas máquinas cometerem erros, tanto quanto é necessário construir uma máquina que seja capaz de explicar, ou indicar de alguma forma, os passos e as informações que foram analisadas para chegar a determinado resultado, pois só assim será possível entender, pelo menos em parte, o comportamento e o processo de tomada de decisão dessas máquinas.

Finalizamos a presente seção reiterando que a possibilidade de cometer erros parece ser um fator fundamental para o desenvolvimento e a adaptação de sistemas inteligentes nos

contextos nos quais estão inseridos. Defendemos ainda que essa capacidade de adaptação dever seria considerada relevante para os sistemas artificiais, assim como é considerada para sistemas naturais, em especial nas etapas de seu desenvolvimento. Assim, em continuidade com a ideia de sistemas que se desenvolvem com o objetivo de adquirir habilidades para poderem agir em determinados ambientes, trataremos na próxima seção sobre a aquisição gradual de autonomia a fim de verificarmos em que medida seria possível responsabilizar um agente artificial por suas ações e escolhas de acordo com seu grau de autossuficiência e independência de controle humano.

3.3 - Graus de autonomia

Na presente seção investigaremos a noção de aquisição gradativa de autonomia e buscaremos explicitar e caracterizar diferentes graus de autonomia a partir dos quais um agente poderia ser, também de acordo com graus, responsabilizado. Pretendemos propor uma escala que permita delimitar graus de autonomia que podem ser adquiridos, partindo da ideia de uma autonomia grau 0 e chegando a uma autonomia que possa ser considerada pragmaticamente plena, caracterizando cada nível de autonomia sugerido.

Nas seções anteriores tratamos de sugestões de autores que defendem a ideia de que certas máquinas têm, ou podem vir a ter, as capacidades necessárias para aprender como uma criança humana. Assim, na primeira seção deste capítulo apresentamos a proposta de Alan Turing (1950) que consiste no desenvolvimento de uma máquina inteligente que tenha as mesmas capacidades de uma criança e que possa vir a desenvolvê-las e ampliá-las por meio de aprendizagem. Já, na segunda seção, apresentamos as ideias de Hauberg, Jørgensen e Detlefsen (2020) que afirmam que uma máquina de inteligência artificial aprende como uma criança, isto é, frequentemente cometendo erros, e que essa disposição da máquina deveria ser explorada a fim de torná-la capaz de reconhecer suas próprias limitações.

Continuaremos a explorar nesta seção a ideia de uma máquina inteligente que se desenvolve como uma criança humana, pois entendemos que o exemplo da criança é eficiente para visualizarmos a aquisição e o desenvolvimento gradativo de uma conduta considerada autônoma e que se uma máquina vier a ser projetada com, pelo menos, alguns princípios de aprendizado humano ela poderá ter mais sucesso em atingir graus elevados de autonomia.

No primeiro capítulo descrevemos os atributos que nos permitem qualificar um agente, tanto natural quanto artificial, como portador de autonomia. De forma resumida é possível dizer que um agente autônomo deve ter a possibilidade de se autorregular, sem que sofra coerções externas, e ter a possibilidade de planificar seu futuro e/ou modificar tal

planejamento. Já para um sistema artificial ser considerado um agente autônomo, ele deve ser capaz de escolher entre opções de acordo com seu contexto e superar sua programação inicial na medida em que aprende com seu ambiente a apresenta *outputs* que não foram previstos por seus programadores.

Entendemos que, tanto no caso de agentes naturais quanto no caso de agentes artificiais, a aquisição de autonomia faz parte de um processo gradativo de desenvolvimento. Isso significa que um agente natural não nasce com os atributos que permitem que ele seja classificado como um agente autônomo: a autonomia será conquistada e desenvolvida na medida em que o agente interage em e com seu ambiente ao longo de sua vida. O mesmo acontecerá com um agente artificial: ainda que um *software* possa ser programado para agir de forma autônoma o programa será mais eficiente, ou mais autônomo, na medida em que tiver mais possibilidades e oportunidades de aprender como ajustar sua conduta a partir de sua programação inicial de modo a alcançar objetivos desejáveis e evitar os indesejáveis sem controle externo.

A fim de caracterizarmos graus de autonomia que possibilitam a consideração de diferentes níveis de responsabilização, iniciaremos nossa investigação com exemplos que ilustrarão algumas fases do desenvolvimento de seres humanos, a maneira como estes adquirem autonomia ao longo de suas vidas e a partir de qual momento passa a fazer sentido atribuir responsabilidade a agentes naturais.

Ressaltamos que devido à complexidade envolvida no processo de desenvolvimento humano, questão ainda em aberto e que é tratada diferentemente por distintas teorias, não caracterizaremos todas as fases de desenvolvimento e consequentemente não referenciaremos todos os graus de autonomia que uma pessoa pode atingir. Porém, acreditamos que é importante mencionar momentos e fases que podem ser consideradas como sendo centrais neste processo para compreendermos com mais clareza o que nos afasta e o que nos aproxima de agentes artificiais, e como podemos compreender o desenvolvimento e a aquisição de autonomia por estes últimos a fim de entendermos se é possível responsabilizá-los por seus atos.

Assim, ao pensarmos em agente humano, é possível considerar que nos primeiros seis meses da vida de um bebê ele não possui autonomia, na medida em que depende totalmente de outras pessoas para poder se locomover, se alimentar, para permanecer limpo, etc. Se pensarmos que a partir dos seis meses de vida um bebê passa a ser capaz de engatinhar, podemos considerar que neste momento ele adquiriu uma primeira parcela de

autonomia, porém ainda depende de um agente externo para que possam realizar funções básicas.

A partir de tentativas e erros uma criança humana, em plenas condições motoras, aprenderá a andar. Neste momento a criança atinge mais uma parcela de autonomia e adquire a capacidade de explorar seu ambiente e suas novas habilidades. Como foi dito anteriormente, entendemos que não faz sentido explorar cada fase do desenvolvimento de uma pessoa, uma vez que não faz sentido responsabilizar crianças muito novas, incapazes de compreender as possíveis consequências de suas ações (não atribuímos responsabilidade a uma criança de 3 anos que, ao mexer no freio de mão de um automóvel, provoque um acidente; responsabilizaremos o adulto que a deixou no interior do carro sem supervisão). Entendemos também que desde muito cedo crianças podem ser capazes de reconhecer quando devem ou não se comportar de determinada maneira na medida em que reconhecem manifestações de aprovação ou desaprovação de seus responsáveis, porém não é possível dizer com certeza quando crianças passam a refletir sobre suas ações e vontades, e quando elas entendem de fato o significado e o alcance de suas intenções e do significado daquilo que seus responsáveis afirmam ser certo ou errado.

Partindo do ponto de que a legislação brasileira, assim como a de muitos países, estabelece que é a partir dos 16 anos de idade que uma pessoa passa a ser capaz de se responsabilizar por certos aspectos de sua vida, parece razoável afirmar que antes disso outra(s) pessoa(s) devem responder legalmente pelos menores de 16, em outras palavras é possível dizer que, de acordo com a lei brasileira, antes dos 16 anos uma pessoa não tem autonomia para exercer atos da vida social como adquirir direitos e contrair deveres. Porém, isso não significa que antes dos 16 anos de idade uma pessoa seja totalmente incapaz de exercer sua autonomia ao praticar certas atividades.

É possível dizer que a cada novo estágio que promove a independência de uma pessoa, sua autonomia está sendo incentivada. Uma criança, ao aprender a andar sozinha, a comer sozinha, a cozinhar sua refeição, a manter seu ambiente organizado e seu corpo limpo, assim como ao manter uma rotina de estudos e administrar seu tempo e suas relações sociais na medida em que constrói sua individualidade e age no mundo, está adquirindo, em cada uma dessas etapas, autonomia para que não dependa de outras pessoas para executar funções que lhe permitam ser considerada um agente autônomo.

Assim sendo, uma pessoa, que neste caso é ainda uma criança ou um adolescente, pode ser responsabilizada por erros ou omissões que de alguma forma tenham impedido ou confundido seu processo de desenvolvimento. E em sentido oposto, porém complementar,

condutas que levam um agente a obter resultados positivos ao tentar atingir um objetivo devem também ser reconhecidas como forma de compensação, dessa forma as ações que tiveram êxito tendem a se repetirem na medida em que podem ser reconhecidas como ações positivas que promovem a conduta acertada em determinadas situações.

Entendemos, assim, que tanto os erros quanto os acertos de um agente devem ser acolhidos e examinados de acordo com cada situação e com cada agente específico, pois responsabilizar um agente apenas por suas falhas pode vir a trazer consequências negativas para o seu desenvolvimento tanto quanto apenas atribuir responsabilidade em forma de premiação e reconhecimento a um agente pelas coisas boas que este conquistou.

É importante enfatizar ainda que para que um agente seja capaz de ajustar sua conduta, e para que seja possível atribuir responsabilidades ao mesmo, os julgamentos acerca de suas ações serão realizados, muitas vezes, por outros agentes, na medida em que cada agente está inserido em ambientes e contextos coletivos.

Assim, é devido à complexidade existente no processo de desenvolvimento do ser humano a partir do qual é possível que adquira sua independência, que encontramos dificuldades em especificar e nomear os graus de autonomia que podem caracterizar este processo, embora possamos vislumbrar etapas no processo contínuo do desenvolvimento. Já no caso de sistemas artificiais que são programados para aprender e agir a partir do mínimo possível de intervenção humana, entendemos que a dificuldade para estabelecer graus de autonomia se amenize devido ao fato de que sistemas artificiais costumam ser elaborados para agir em ambientes específicos. Isso significa que, ainda que não compreendamos toda a programação e o processamento algorítmico que constituem um sistema artificial, é possível considerar que sua ação, assim como sua interação, não sejam tão modificáveis por fatores externos, pelo menos no caso de sistemas artificiais inteligentes cujo objetivo é executar tarefas determinadas.

Isto posto, passamos a examinar os *outputs* e o alcance de certos artefatos tecnológicos construídos para exibirem tanto comportamento inteligente quanto algum tipo de conduta autônoma. Partiremos do grau zero, no qual um sistema é inteiramente dependente de controle humano para que possa cumprir com suas funções e chegaremos ao grau quatro, no qual um sistema seria capaz de agir ou tomar decisões sem supervisão, aprovação ou desaprovação externa.

Grau zero de autonomia - nesta categoria estão inclusos sistemas que podem ser considerados especialistas, mas que precisam de intervenção humana para poderem funcionar. Como exemplos de aparatos tecnológicos capazes de instanciar funções inteligentes, porém

sem autonomia, é possível citar calculadoras, máquinas de lavar conectadas à rede de internet, cafeteiras programáveis, relógios, tecnologias de reconhecimento de voz, entre outros. Nestes casos se o sistema apresentar alguma falha, a responsabilidade pela mesma será atribuída ao mau uso das tecnologias ou à uma possível fabricação defeituosa.

Grau um de autonomia - nesta categoria é possível incluir artefatos com *software* os quais são capazes de aprender e filtrar informações a partir da interação em seu ambiente e com outros usuários do sistema, por exemplo: assistentes virtuais digitais, *chatbots*, robôs aspiradores, jogos *online* como xadrez e *poker*, ou outras plataformas que forneçam as condições para que um usuário possa interagir com um programa de computador que cumpre o papel de outro jogador.

Entendemos que tecnologias desta categoria são frequentemente programadas para aprender a agir de acordo com a conduta de um usuário específico, ou seja, todo o *output* que a tecnologia exibir será condizente com o comportamento daquele que a utiliza com mais frequência: por exemplo, assistentes virtuais captam, coletam e assimilam os dados do usuário, os mantêm em sua memória, e agem, ou provêm respostas de acordo com preferências comuns anteriores deste usuário. *Chatbots* costumam ter respostas pré-programadas que dependerão de perguntas externas. Robôs aspiradores são capazes de mapear o ambiente no qual estão inseridos e aspirar de acordo com as preferências e programação de seus usuários. Programas de jogos *online* são elaborados para coletar e armazenar incontáveis informações sobre jogadas e regras que lhes permitem realizar cálculos para jogadas futuras.

Em todos os casos anteriores, as tecnologias também não são passíveis de responsabilização uma vez que, ainda que apresentem algum grau de autonomia, elas dependem de uma pessoa para que possam realizar aquilo para o qual foram programadas, assim, se essas tecnologias, como as tecnologias de grau zero, são suscetíveis a falhas técnicas, porém não a erros resultantes de escolha deliberada.

Grau dois de autonomia - incluimos nesta categoria tecnologias de piloto automático que venham a possibilitar que alguns veículos se locomovam sem supervisão humana, ou com o controle mínimo de um agente humano. O piloto automático é um serviço utilizado em submarinos, aviões, carros, veículos espaciais, trens e metrô e possibilita que os pilotos desses veículos não precisem ficar atentos a todos os detalhes de sua viagem; em alguns casos, o piloto automático é capaz de conduzir um veículo em todo o trajeto que deve ser percorrido. No caso de aviões, de acordo com Alisson Hope (2017):

O piloto automático se refere a uma coleção de sistemas que automatizam as operações de um avião. A complexa matriz do computador informa ao seu avião como voar, incluindo navegação, altitude, velocidade e impulso do motor, que controla a força pela qual o avião se move no ar. Quando esses sistemas estão ativados - depois que um humano insere as informações de destino do voo, o piloto automático seleciona dados sobre a rota de voo, localização e navegação - o sistema de navegação usa a mesma tecnologia GPS que está em seu telefone celular e lança um plano de voo otimizado. Isso permite que o piloto permaneça com as mãos livres durante o voo. (Tradução nossa)⁷⁶

Enfatizamos assim que o piloto automático de um carro apresenta grandes diferenças de um piloto automático de um avião ou de um trem: em carros, o piloto automático permite que o condutor do mesmo fixe uma velocidade específica, mas não permite que o carro se guie sozinho. Já no caso de aviões ou trens, o caminho a ser percorrido é estabelecido antes que a viagem seja iniciada, assim, de acordo com a programação e a capacidade de cada veículo, a tecnologia de piloto automático possibilita que seus condutores realizem outras atividades na medida em que não necessita ser conduzido por um agente humano.

Entendemos que tecnologias de piloto automático dependem de uma série de variações externas para que seu funcionamento ocorra da melhor forma possível, ou seja, para que não ocorram acidentes ou para que caso ocorram, que as consequências possam ser amenizadas e contidas de alguma forma, neste sentido é necessário que esses sistemas sejam construídos com a capacidade de identificar possíveis riscos e agir diante desses riscos.

Consideramos assim que, tecnologias de piloto automático podem ser incluídas na categoria de agentes que podem vir a ser penalizados por seus erros ou por suas falhas em um sentido minimalista: se o *software* causou um dano significativo decorrente de seu mau funcionamento, esse programa poderá ter sua utilização proibida por agências reguladoras.

Desse modo, embora possa ser atribuída ao *software* corresponsabilidade (juntamente com seus idealizadores, distribuidores e usuários) pelos danos causados, até por se tratar de um sistema especialista, a despeito de poder ser considerado autônomo, não parece apropriado considerar tal sistema um agente moral. Mas entendemos que em casos nos quais pessoas passam a depender do bom funcionamento de uma máquina com algum grau de

⁷⁶ Autopilot refers to a collection of systems that automate a plane's operations. The complex computer matrix tells your plane how to fly, including navigation, altitude, speed, and engine thrust, which controls the force by which the plane moves through the air. When these systems are engaged—after a human enters the flight destination information, autopilot culls data about the flight route, location, and navigation—the navigation harnesses the same GPS technology that's on your cell phone and spits out an optimized flight plan. This allows the pilot to remain hands-free for the duration of the flight.

autonomia, é necessário incluir essas máquinas na rede de agentes corresponsáveis pela ação em questão.

Grau três de autonomia - nesta categoria incluímos programas que sejam capazes de buscar informação, processá-la e realizar previsões e, especialmente, por decisão humana de seus idealizadores e usuários, lhes é conferido o poder de tomar decisões sem supervisão de um agente humano. Estes programas não precisam necessariamente estar incorporados em corpos robóticos ou em algum tipo de *hardware* como nos exemplos de tecnologias que possuem grau dois de autonomia.

Como exemplo de tecnologias com grau três de autonomia é possível citar o caso da multinacional *Amazon* que vem sendo acusada por empregar em suas unidades um *software* para vigiar constantemente seus funcionários. A tecnologia utilizada pela *Amazon* é capaz de monitorar o desempenho de seus funcionários, medindo a produtividade e o período de pausa dos mesmos. Como se a invasão da privacidade dos funcionários e a desconsideração pela capacidade individual de trabalho de cada um não fossem o suficiente, em algumas unidades da empresa, essas tecnologias contam ainda com a capacidade e a autorização para emitir alertas e avisos sobre a baixa produtividade de funcionários e até demiti-los caso metas de produção determinadas pela empresa não sejam cumpridas.

Entendemos que este tipo de aplicação de tecnologia inteligente pode ser considerado como um evento gravíssimo que mostra como a tecnologia pode interferir negativamente na conduta e na autonomia das pessoas, afetando seu bem-estar e seu desenvolvimento interpessoal.

Consideramos ainda que, embora a este tipo de tecnologia tenha sido atribuída a autonomia para interferir diretamente na vida de um ser humano, a responsabilidade por suas decisões deve ainda ser atribuída àqueles que autorizam o seu uso e que determinam a partir de quais informações o sistema funcionará. Assim, em casos como o da *Amazon*, o processo de demissão e de punição gerado pela máquina deveria passar por escrutínio de supervisores humanos, que deveriam ser capazes de analisar cada caso em particular e decidir se de fato uma pessoa deve ou não continuar em seu posto de trabalho, assim como devem ponderar e decidir se as advertências dadas pela máquina são pertinentes ou não.

Grau quatro de autonomia - trataremos a presente categoria como o último grau de autonomia que uma tecnologia pode vir a adquirir a partir de sua programação e de seu desenvolvimento enquanto agente, ou seja, nesta categoria espera-se que os agentes artificiais, após sua programação inicial e após o fornecimento de *inputs* básicos, sejam capazes de interagir com seu ambiente na medida em que captam e filtram informações e buscam,

sozinhos, soluções para a resolução de suas tarefas ultrapassando os limites algorítmicos de sua programação inicial.

Como exemplos de tecnologias que alcançariam o grau quatro de autonomia citamos os projetos de carros autônomos e os robôs cuidadores, ainda em desenvolvimento. Como dito anteriormente, o que se espera de um carro autônomo inteligente é que este seja capaz de conduzir passageiros do ponto A ao ponto B sem que o passageiro tenha que se preocupar com detalhes do trânsito ou do próprio veículo, ou seja, espera-se que a pessoa seja apenas um passageiro e não o condutor do veículo. Assim, bastaria que a pessoa escolha seu destino e o carro deverá calcular e decidir qual o melhor caminho a ser percorrido de acordo com as necessidades de seu passageiro e as informações contextuais que é capaz de coletar e processar. E em casos de situações de possível risco aos passageiros do carro, a outros veículos, pedestres e animais ao entorno do veículo autônomo, este deve ser capaz de encontrar uma solução que tenha como objetivo abrandar as possíveis consequências negativas e os danos que podem decorrer da situação.

Um robô cuidador, por sua vez, deve ser capaz de encarregar-se inteiramente pelo bem-estar da pessoa sob seu cuidado. O seu grau de autonomia em relação aos cuidados da pessoa deverá ser proporcional à perda de autonomia da pessoa. É possível considerar que um robô cuidador seja elaborado para substituir profissionais como babás e enfermeiros. Assim, é necessário que o robô seja construído para ser capaz de realizar tanto atividades básicas quanto atividades complexas que são esperadas desses profissionais.

Entendemos ainda que, um robô cuidador pensado para executar a função de um profissional como uma babá é diferente de um robô cuidador pensado para substituir um profissional da enfermagem. Assim sendo, cada sistema será programado de acordo com sua função específica, e ainda que este tipo de robô deva ser programado para aprender e para agir de maneira autônoma, é necessário enfatizar que mesmo robôs inteligentes e autônomos são pensados para agir em seus contextos específicos, assim é necessário que o robô receba e assimile o maior número de informações possíveis sobre o ambiente no qual atuará, para que seja capaz de se locomover sem causar acidentes ou ficar preso em alguma situação, tanto quanto deve receber o maior número possível de informações sobre as pessoas sob seu cuidado, para que seja capaz, por exemplo, de administrar remédios, monitorar o estado de saúde e perceber quando há algo errado ou anômalo.

Cabe enfatizar que embora existam robôs que cumprem funções de assistentes e carros com grandes capacidades automatizadas, robôs cuidadores e carros autônomos como descritos acima ainda não fazem parte da realidade das pessoas. Devido ao elevado grau de

autonomia que se espera dessas máquinas, é necessário muito estudo acerca da programação que possibilitará que esses sistemas se comportem efetivamente como agentes, além de entendimento de como essas ações desses sistemas podem afetar aqueles que deles dependem.

Entendemos assim que seria possível e até mesmo desejável incluir esses agentes artificiais na categoria de agentes responsabilizáveis.

A fim de concluir o presente capítulo, destacamos novamente que a intenção em se buscar inserir agentes artificiais na categoria de agentes que podem ser responsabilizados e, conseqüentemente, poderem ser alvo de algum tipo de punição ou recompensa, no sentido proposto por Turing (1950), não implica a transferência de responsabilidade de agentes naturais para agentes artificiais, ou seja, não admitimos, em momento algum, a exclusão ou omissão dos agentes naturais que fizeram parte de todo o processo de desenvolvimento e construção de um sistema artificial bem como os agentes que fazem uso desses sistemas das redes de responsabilização que permeiam processos complexos.

Teceremos a seguir nossas considerações finais a fim de retomarmos o caminho feito ao longo do texto que nos permitiu considerar que sistemas artificiais podem ser inseridos na categoria de agentes autônomos passíveis de atribuição de responsabilidade, sem necessariamente constituírem agentes morais no sentido pleno, e trataremos ainda do porquê é necessário pensar nesses sistemas enquanto agentes responsabilizáveis e como é possível responsabilizá-los.

CONSIDERAÇÕES FINAIS

O objetivo central desta dissertação foi investigar a possibilidade de sistemas artificiais autônomos poderem ser considerados agentes morais aos quais cabe atribuir algum grau significativo de responsabilidade.

A fim de realizar nosso objetivo, percorremos, ao longo de três capítulos, um caminho que nos permitiu indicar como a presença de sistemas artificiais, considerados inteligentes e autônomos em algum grau, pode impactar a vida dos seres humanos e porque é pertinente pensar nesses sistemas enquanto agentes com a capacidade de gerar consequências a partir de suas ações, assim como de influenciar e intervir na tomada de decisão de agentes naturais.

Tendo em vista o poder de ação de certos sistemas artificiais, indicamos que é importante pensar nesses sistemas enquanto agentes que podem ser responsabilizados por suas ações, e ainda que, por se tratarem de sistemas especialistas para a realização de funções pré-estabelecidas, não possam ser considerados agentes morais no mesmo sentido como são considerados os seres humanos, entendemos que agentes artificiais estão envolvidos em situações morais na medida em que podem, a partir de suas ações, gerar consequências desejáveis ou indesejáveis.

Entendemos que os processos de fabricação e uso de tecnologias altamente sofisticadas, como as que tratamos nesta pesquisa, envolve uma grande rede de pessoas e essa condição central pode dificultar o processo de atribuição de responsabilidades quando a tecnologia em questão causa algum tipo de dano ou prejuízo. Porém enfatizamos que a nossa investigação, que busca inserir sistemas artificiais autônomos na categoria de agentes que podem ser responsabilizados, não pretende em momento algum transferir a responsabilidade dos agentes naturais aos agentes artificiais, mas sim incluir os últimos na rede de agentes encarregados de cumprir alguma função que pode gerar consequências positivas ou negativas e estariam, assim, sujeitos a receberem algum tipo de recompensa ou punição, como sugerido por Turing (1950).

Para chegar a essa conclusão esta pesquisa estruturou-se da seguinte forma: no *Capítulo 1 - Análise do Conceito de Agência*, foi elaborado uma análise sobre o conceito de **agência** e sobre o conceito de **autonomia** a fim de verificarmos se é justificável atribuir as características de um agente autônomo a um sistema artificial.

Verificamos que para Anscombe (1975) e para Davidson (1963), autores contemporâneos da Filosofia da Mente e da Filosofia da Ação, a noção de agência é

sustentada pela noção de ação intencional, isso significa que apenas organismos com capacidades cognitivas de segunda ordem seriam capazes de ser considerados agentes no sentido rigoroso do conceito.

Porém, ao seguirmos com a investigação verificamos que existem tipos diferentes de agência que não requerem a atribuição de representações mentais (Schlosser, 2019). Isso significa que organismos que não possuam necessariamente capacidades cognitivas ditas de segunda ordem, como fazem os agentes humanos, podem também possuir algum tipo legítimo de agência, isto posto, caracterizamos as noções de agência mental, agência epistêmica, agência compartilhada, agência coletiva, agência relacional e, por fim, a noção de agência artificial.

O passo seguinte do nosso trabalho consistiu na caracterização do conceito de autonomia, uma vez que entendemos que a autonomia é uma característica estrutural que permite pensar a noção de agência de maneira não trivial.

Concluimos o primeiro capítulo com a investigação que nos permitiu verificar que sistemas artificiais podem ser considerados agentes portadores de autonomia em algum grau. É importante ressaltar que nem todas as características de um agente natural autônomo se aplicam a um agente artificial autônomo, contudo isso não exclui necessariamente os agentes artificiais da categoria de agentes autônomos.

No *Capítulo 2 - Conduta Moral e Tipos de Responsabilidade* introduzimos o filósofo Hans Jonas (2006) que nos permitiu tratar sobre o tema da atribuição de responsabilidade e verificar que esta nem sempre estará relacionada a um sentimento moral.

Tratamos sobre as noções de responsabilidade coletiva e responsabilidade compartilhada, o que propiciou uma discussão acerca de ações praticadas por mais de um agente e como é possível tratar da atribuição de responsabilidade em casos de ações coletivas.

Analizamos ainda como as tecnologias de informação e comunicação (TIC) vem influenciando a rotina das pessoas e modificando suas ações e a forma como se relacionam. Destacamos que, embora discussões filosóficas acerca da atribuição de responsabilidade moral tiveram como foco a agência humana, na sociedade tecnológica atual tratar sobre a agência humana sem fazer referências a sistemas tecnológicos é um desafio, e que essa circunstância pode dificultar a atribuição de responsabilidade, uma vez que a realidade da computação ubíqua compreendida a partir uso de aparatos tecnológicos conectados continuamente à rede de internet distancia o agente dos desdobramentos e das consequências de suas ações.

Após uma análise acerca da atribuição de responsabilidade e da conduta responsável de agentes naturais em concomitância com o uso das TIC, finalizamos o segundo capítulo com a investigação sobre a possibilidade de incluir agentes artificiais portadores de autonomia na categoria de agentes responsabilizáveis, uma vez que essas tecnologia vem sendo desenvolvidas e aplicadas em contextos causais nos quais, numa realidade pré tecnologias autônomas, apenas seres humanos seriam capazes de praticar ações significativas.

Entendemos que agentes artificiais são suscetíveis à responsabilização na medida em são agentes portadores de algum grau de autonomia que fazem parte de contextos que dependem de seu bom funcionamento para que uma situação possa acontecer como foi planejada sem acidentes ou prejuízo àqueles que dela fazem parte.

No *Capítulo 3 - Máquina Criança e Graus de Autonomia* exploramos a ideia de Alan Turing (1950) sobre a construção de uma máquina inteligente com as capacidades de aprendizado de uma criança. Turing (1950) propôs que se uma máquina viesse a ser elaborada com as mesmas capacidades de desenvolvimento de uma criança, esta poderia atingir as capacidades mentais de um adulto humano.

Em conjunto com a proposta de Turing, examinamos a ideia de que máquinas inteligentes devem ser programadas para ser capazes de cometer erros e ajustar sua conduta na medida em que aprendem com esses erros. Defendemos que o erro é um importante aspecto do desenvolvimento e da adaptação de um sistema, seja ele natural ou artificial, e que, assim sendo, se pretendemos permitir que máquinas inteligentes façam parte da rotina humana tomando decisões importantes e praticando ações significativas (e que podem causar danos ou benefícios a pessoas), é fundamental considerar que essas máquinas devem ser programadas para ter a capacidade de errar e de reconhecer seus erros.

Ao nos apoiarmos nas ideias de uma máquina que atinja certa maturidade, no sentido de Turing (1950) e independência após ter sido construída para se desenvolver e aprender como uma criança humana, concluímos o terceiro capítulo caracterizando graus de autonomia a partir dos quais seria possível compreender a aquisição de autonomia por parte de uma máquina e a atribuição de responsabilidade em graus.

Percorremos um caminho que nos levou do grau zero de autonomia, no qual dispositivos tecnológicos são totalmente dependentes de fatores externos para que possam cumprir suas funções, e chegamos ao grau quatro de autonomia, no qual sistemas artificiais de tecnologia avançada teriam capacidade de ação e de tomar decisões, não dependendo mais de seres humanos para alimentá-los com *inputs* na espera de resultados e decisões previsíveis.

Tecemos assim, ao longo de três capítulos, uma investigação que nos permitiu concluir que sistemas artificiais podem ser considerados agentes tanto quanto podem ser considerados sistemas autônomos. A partir dessa investigação entendemos que esses agentes artificiais autônomos estão sendo elaborados para, de maneira significativa, fazerem parte do cotidiano e da rotina dos seres humanos.

A esses sistemas altamente sofisticados, que tendem a ser aprimorados com o tempo, serão reservadas tarefas e funções que antes só poderiam ser realizadas por seres humanos, e este fato é uma das principais motivações que nos levam a pensar nesses sistemas enquanto agentes que podem ser corresponsabilizados na medida em que têm capacidades de ação com desdobramentos morais.

Partindo do pressuposto de que sistemas artificiais autônomos especialistas não possuem capacidades análogas às dos seres humanos que justificariam sua responsabilização exclusiva, assim como não possuem a capacidade de processar emoções, tais como culpa ou orgulho, entendemos que ainda não seja possível atribuir responsabilidade a sistemas artificiais da mesma forma que a responsabilidade é atribuída aos seres humanos, embora seja possível conceber inputs de recompensa ou punição que, como sugere Turing (1950), auxiliariam no processo de aprendizagem de máquinas criança.

Assim, quando um sistema artificial falha de alguma maneira e essa ação ocasiona danos e prejuízos, é necessário que seja possível investigar a origem dessa falha a fim de verificar a rede de corresponsabilização pelos danos causados. Elementos de tal rede podem vir a ser, a princípio, a empresa responsável pelo sistema ou o(s) usuário(s) do sistema em questão. Ao sistema ficaria reservada uma interdição, assim como uma correção de programação, até que seja certificada sua capacidade de ser utilizado novamente com segurança.

Supomos também que, se sistemas artificiais forem construídos com certas capacidades que proporcionem algum tipo de “entendimento” (cujas natureza, porém, ignorados), eles serão capazes de perceber o significado de um erro e de um acerto, de uma ação que teve sucesso e de uma ação que não atingiu seu objetivo. Entendemos que essas habilidades permitirão que um sistema artificial adeque sempre sua conduta na medida em que tenta cumprir suas funções sem causar danos a terceiros.

Porém, enquanto isso ainda não acontece, acreditamos ser necessário continuar pensando esses sistemas enquanto agentes autônomos que cumprem um papel significativo em uma rede de ações que agrega diversos fatores e diferentes agentes, e que as ações praticadas por agentes artificiais autônomos, mas não necessariamente agentes morais plenos,

podem produzir tanto resultados desejáveis quanto resultados indesejáveis, tenham sido eles previstos ou não.

REFERÊNCIAS

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

BARANDIARAN, X. E, DI PAOLO E., ROHDE M.: Defining Agency: individuality, normativity, asymmetry, and spatio-temporality in action, *Adaptive Behavior*, v. 17.n. 5, p. 367–386, 2009.

BEAVERS, Anthony. F. Moral machines and the threat of ethical nihilism. In: Lin, P.; Bekey, G. & Abney, K. (Ed.). *Robot ethics: the ethical and social implications of robotics*. Cambridge: The MIT Press, p. 333-44, 2011.

BRASIL. Código Civil. Disponível em:

<http://www.planalto.gov.br/ccivil_03/leis/2002/110406compilada.htm>. Acesso em: 12. nov. 2020.

BRASIL. Código Penal. Disponível em:<http://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm>.Acesso em 06 nov. 2020.

BROENS, Mariana C. Self-Organization and action: a systemic approach to common action. In: PEREIRA JR. Alfredo; PICKERING, WILLIAM A. & GUDWIN, Ricardo R. *Systems, Self-organization and Information: An interdisciplinary perspective*. New York: Routledge, p. 127-138, 2018.

BUSS, Sarah.; WESTLUND, Andrea. *Personal Autonomy*, The Stanford Encyclopedia of Philosophy. Edward N. Zalta (ed.) Disponível em: <<https://plato.stanford.edu/archives/spr2018/entries/personal-autonomy/>>.Acesso em :06 nov. 2020.

COMISSÃO INTERNACIONAL PARA QUALIFICAÇÃO DE TESTE DE SOFTWARE. *Base de Conhecimento para Certificação em Teste*, 2018.

CLARKE, Arthur, C. 2001: uma odisseia no espaço. São Paulo: Aleph, 2013.

DAVIDSON, Donald. *Actions, Reasons, and Causes* (1963) in *Essays on Actions and Events*. 2. ed. New York: Oxford Clarendon Press, 2002. p. 3 - 19.

DAVIDSON, Donald. *Agency (1971) in Essays on Actions and Events*. 2.ed. New York: Oxford University,p. 43 – 61, 2002.

DENNETT, Daniel C. *When hal kills, who's to blame?: computer ethics*. Cambridge, MA: MIT Press, 1997. p. 351-365. Disponível em: < <http://hdl.handle.net/10427/57613>> Acesso em: 03 nov.2020.

DRIVER, Julia, Gertrude Elizabeth Margaret Anscombe, *The Stanford Encyclopedia of Philosophy*. Disponível em: <<https://plato.stanford.edu/archives/spr2018/entries/anscombe/>>.Acesso em: 10 jun.2020.

DRYDEN, Jane. *Autonomy*. The Internet Encyclopedia of Philosophy. Disponível em: <<https://iep.utm.edu/autonomy/>> Acesso em:17 jun. 2020.

EUROPEAN COMMISSION. *Artificial intelligence, robotics and 'autonomous' systems*. Luxembourg: publications office of the European Union, 2018. Disponível em: <https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf>. Acesso em: 15 out 2018.

FLORIDI, Luciano. SANDERS, J.W. On the morality of artificial agents. *Minds and Machines*. Netherlands. *Kluwer Academic Publishers*, v. 14 p. 349-379, 2004.

FRANKLIN, S.; GRAESSER, A. *Is it an agent, or just a program?* a taxonomy for autonomous agents. Institute for Intelligent Systems - University of Memphis. Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages: Springer-Verlag, 1996.

GERT, Bernard; GERT, Joshua. The definition of morality. *The Stanford Encyclopedia of Philosophy*. Disponível em: <<https://plato.stanford.edu/archives/fall2017/entries/morality-definition/>>. Acesso em: 10 mai.2020.

GOGONI, Ronaldo. *O que é software?* Tecnoblog. Disponível em: <<https://tecnoblog.net/311647/o-que-e-software/>>. Acesso em: 07 out. 2020.

GONZALEZ, Maria. E. Q. Autonomous action in complex mechanical systems: a real dilemma?. In: Frederick Adams; Osvaldo Pessoa Jr.; João Eduardo Kogler Jr. (Org.). *Cognitive Science: Recent advances and recurrent problems*, v. 1, p. 17-30. Delaware: Vernon Press, 2017.

GONZALEZ, Maria. E. Q.; BROENS, Mariana. C.; MORAES, João. A. A virada informacional na filosofia: alguma novidade no estudo da mente? In: *Aurora*. Curitiba: PUCPR, v. 22. n. 30, p. 137-151, jan/jul, 2010.

HALDANE, John. In memoriam: G. E. M. Anscombe (1919-2001). *The Review of Metaphysics*. v. 53 n.4, p. 1019-1021, 2000. Disponível em: <<https://www.st-andrews.ac.uk/~jjh1/Haldob.pdf>>. Acesso em: 7 jun. 2020.

HAUBERG, Soren. JORGENSEN, Martin. DETLEFSEN, Nicki. S. Computers must learn to make mistakes. *ScienceNordic*, 2020. Disponível em: <<https://sciencenordic.com/denmark-technology/computers-must-learn-to-make-mistakes/1678962>>. Acesso em: 09 out. 2020.

HAUBERG, Soren. JORGENSEN, Martin. DETLEFSEN, Nicki. S. *Reliable Training and Estimation of Variance Networks*. 33. ed. Conference on Neural Information Systems (NeurIPS 2019), Canadá : Vancouver, 2019.

HOPE, Allison. *How autopilot on planes works*. Condé Nast Traveler, 2017. Disponível em: <<https://www.cntraveler.com/story/how-autopilot-on-planes-works>>. Acesso em 05 nov. 2020.

JACOB, Pierre. *Intentionality*. Edward N. Zalta (ed.). 2014. Disponível em: <<https://plato.stanford.edu/archives/win2014/entries/intentionality/>>. Acesso em: 23 out.2020.

JEE, Charlotte. *Amazon's system for tracking its warehouse workers can automatically fire them*. MIT Technology Review. Disponível em:<<https://www.technologyreview.com/2019/04/26/1021/amazons-system-for-tracking-its-warehouse-workers-can-automatically-fire-them/>>. Acesso em: 20 out. 2020.

JONAS, Hans. *O princípio responsabilidade: ensaio de uma ética para a civilização tecnológica*. Rio de Janeiro: PUC-Rio, 2006.

JOY, Bill. Why the future doesn't need us, *Wired Magazine*. 2000. Disponível em: <<https://www.wired.com/2000/04/joy-2/>> Acesso em: 10 out. 2019.

KURZWEIL, Raymond. *A lei do tempo e do caos: a era das máquinas*. Trad. Fábio Fernandes. p. 27-66, 2007.

MALPAS, Jeff. Donald Davidson. *The Stanford Encyclopedia of Philosophy*. Disponível em:<<https://plato.stanford.edu/archives/fall2019/entries/davidson/>>. Acesso em 19 jun.2019.

MATHIAS, Andreas. The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* v. 6, p. 175–183, 2004.

MILLER, Seumas. collective responsibility and information and communication technology. In: *Information Technology and Moral Philosophy United States of America*: Cambridge University Press, p. 226 – 250, 2008.

NISSENBAUM, Helen. *Accountability in a computerized society*. Science and Engineering v. 2 p. 25-4, 1996.

NOORMAN, Merel, Computing and moral responsibility. The Stanford Encyclopedia of Philosophy, 2020. Disponível em :<<https://plato.stanford.edu/archives/spr2020/entries/computing-responsibility/>>. Acesso em: 15 ago.2020.

O ROBÔ DA AMAZON QUE DEMITE OS FUNCIONÁRIOS. Instituto Humanas Unisinos, 2019. Disponível em:<<http://www.ihu.unisinos.br/78-noticias/588859-o-robo-da-amazon-que-demite-os-funcionarios>> Acesso em: 11 out. 2020.

PITT, David. Mental representation. *The Stanford Encyclopedia of Philosophy*, 2020. Disponível em :<<https://plato.stanford.edu/archives/spr2020/entries/mental-representation/>>.Acesso em: 20 set.2020.

QUILICI-GONZALEZ, José Artur et al. Complexity and information technologies: an ethical inquiry into human autonomous action. *Sci. stud.* São Paulo, v. 12, p. 161-179, 2014.

RICHTER, Duncan. G. E. M. Anscombe (1919-2001). *The Internet Encyclopedia of Philosophy*, p. 2161-0002. Disponível em: <<https://www.iep.utm.edu/anscombe/#H3>>. Acesso em: 7 jun. 2020.

SCHLOSSER, Markus, Agency. *The Stanford Encyclopedia of Philosophy*, 2019. Disponível em:<<https://plato.stanford.edu/archives/win2019/entries/agency/>>.Acesso em: 10 out. 2020.

SEARLE, John. *Mente, Cérebro e ciência*. Lisboa: Edições 70. 1984.

SMILEY, Marion, Collective responsibility, *The Stanford Encyclopedia of Philosophy*, 2017. Disponível em :<<https://plato.stanford.edu/archives/sum2017/entries/collective-responsibility/>>. Acesso em: 22 out.2020.

SULLINS, John, Information technology and moral values. *The Stanford Encyclopedia of Philosophy*, 2019, Disponível em: <<https://plato.stanford.edu/archives/sum2019/entries/it-moral-values/>>.Acesso em : 22 nov.2020.

SULLINS, John P. When is a robot a moral agent? *International Review of Information Ethics*, v. 6, n. 12, p. 23-30, 2006.

TEIXEIRA, João F. de. *Inteligência artificial*. São Paulo: Paulus. 2009

TEIXEIRA, João F. de. *O cérebro e o robô: inteligência artificial, biotecnologia e a nova ética*. São Paulo: Paulus, 2015.

TEIXEIRA, João F. de. *O pesadelo de Descartes: do mundo mecânico à inteligência artificial*. [Recurso eletrônico]. Porto Alegre, RS: Editora Fi, 2018. Disponível em: <<https://www.editorafi.org/405joaoteixeira>>. Acesso em: 11 set. 2019.

THE ROYAL ACADEMY OF ENGINEERING. *Autonomous Systems: social, legal and ethical issues*. London, 2009. Disponível em: <<https://www.raeng.org.uk/news/news-releases/2009/August/autonomous-systems-social-legal-and-ethical-issues>>. Acesso em: 19 abr. 2019.

TURING, Alan. Computing machinery and intelligence. In: VITZ, Rico. Doxastic Voluntarism. *The Internet Encyclopedia of Philosophy*, v. 54, p. 236-45, 1950. Disponível em: <<https://www.iep.utm.edu/doxa-vol/#H3>> Acesso em:12 jun. 2020.

WEBER, Marc. *Where to?* A history of autonomous vehicles. CHM Blog, 2014. Disponível em: < <https://computerhistory.org/blog/where-to-a-history-of-autonomous-vehicles/>>. Acesso em 07 out. 2020.

WEISER, Mark. The Computer for the 21st Century. In *Scientific American*, v. 265, n. 3, p. 94–105,1991.

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