UNIVERSIDADE FEDERAL DE PELOTAS
INSTITUTO DE FILOSOFIA, SOCIOLOGIA E POLÍTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA

Tese de Doutorado

Sobre a viabilidade da representação de conhecimento moral em sistemas inteligentes: uma abordagem baseada em lógica não-monotônica.

Flávia Braga de Azambuja

Pelotas, 2024

Flávia Braga de Azambuja

Sobre a viabilidade da representação de conhecimento moral em sistemas inteligentes: uma abordagem baseada em lógica não-monotônica.

Tese apresentada ao Programa de Pós-Graduação em Filosofia do Instituto de Filosofia, Sociologia e Política da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Doutor em Filosofia. Área de concentração: Epistemologia Moral

Orientador(a): Juliano Santos do Carmo

Pelotas, 2024

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação da Publicação

A991s Azambuja, Flávia Braga de

Sobre a viabilidade da representação de conhecimento moral em sistemas inteligentes [recurso eletrônico] : uma abordagem baseada em lógica não-monotônica / Flávia Braga de Azambuja ; Juliano Santos do Carmo, orientador. — Pelotas, 2024.
118 f.

Tese (Doutorado) — Programa de Pós-Graduação em Filosofia, Instituto de Filosofia, Sociologia e Política, Universidade Federal de Pelotas, 2024.

1. Conhecimento moral. 2. Sistemas inteligentes. 3. Lógica. I. Carmo, Juliano Santos do, orient. II. Título.

CDD 100

Elaborada por Michele Lavadouro da Silva CRB: 10/2502

Flávia Braga de Azambuja

Sobre a viabilidade da representação de conhecimento moral em sistemas inteligentes: uma abordagem baseada em lógica não-monotônica.

Tese apresentada, como requisito parcial para obtenção do grau de Doutor em Filosofia, Programa de Pós-Graduação em Filosofia, Instituto de Filosofia, Sociologia e Política, Universidade Federal de Pelotas.

Data da defesa: 05/07/2024

Banca examinadora:

Prof. Dr. Juliano Santos do Carmo (Orientador)
Doutor em Filosofia pela Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

Prof. Dr. Prof. Dr. Evandro Barbosa
Doutor em Filosofia pela Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

Prof. Dr. Carlos Alberto Miraglia
Doutor em Filosofia pela Universidade Federal de Pelotas

Prof. Dr. Anderson Priebe Ferrugem
Doutor em Ciência da Computação pela Universidade Federal de Pelotas

Prof. Dr. Dante Augusto Couto Barone
Doutor em Informática pelo Institut National Polytechnique de Grenoble (INPG)

Agradecimentos

Agradeço primeiramente a Deus, por me dar saúde e serenidade para superar os desafios e concluir este trabalho.

Um imenso agradecimento ao meu companheiro de vida Gustavo Zechlinski, cujo apoio, paciência e amor foram fundamentais em cada página escrita. Você foi meu porto seguro e minha inspiração diária.

Aos meus queridos filhos, que me lembraram sempre da alegria e do propósito maior de minha jornada. Obrigado por entenderem os momentos de dificuldade, que foram compensados com a esperança de um futuro melhor para todos nós. Um agradecimento especial ao João Vitor Carvalho e à Morgana Mesquita pela ajuda na revisão e formatação.

Um sincero agradecimento ao meu orientador, cuja sabedoria, dedicação e rigor acadêmico me guiaram nesta jornada intelectual. Sua orientação foi essencial para o desenvolvimento e aprimoramento deste trabalho.

Um agradecimento especial a mim mesma, por não desistir de mim, por insistir, persistir, superar e reunir toda a minha coragem para finalizar este ciclo.

Tenho morrido muitas vezes.
Depois, respiro fundo, lavo o rosto, sigo em frente.
Não é fácil morrer, difícil é renascer, fingir-se de
sol, cegar a lua, beber o mar. Detestável seria ter
a covardia dos que me mataram. Eu sigo
renascendo, eles seguem covardes.
(Pedro Munhoz)

Resumo

AZAMBUJA, Flávia Braga de. **Sobre a viabilidade da representação de conhecimento moral em sistemas inteligentes: uma abordagem baseada em lógica não-monotônica.** Orientador: Juliano Santos do Carmo. 2024. 114 f. Tese (Doutorado em Filosofia) – Instituto de Filosofia, Sociologia e Política, Universidade Federal de Pelotas.

Esta tese investiga a capacidade de sistemas de inteligência artificial (IA) para engajar-se em tomadas de decisão moralmente significativas. A hipótese geral postula que, embora a IA esteja se tornando onipresente em contextos que exigem julgamentos morais, não podemos atribuir responsabilidade moral a sistemas autônomos, pois o conhecimento moral não pode ser completamente representado ou simulado por lógica. As hipóteses específicas exploram várias nuances dessa questão central. Argumenta-se que sistemas autônomos não possuem "motivação moral" e que, apesar da capacidade da IA de automatizar alguns processos de pensamento humano, como criatividade e decisões, ela não pode gerar conhecimento moral autêntico devido ao impacto do viés de dados e à natureza condicional e contextual do conhecimento. A tese também sustenta que não é possível estabelecer uma base para o conhecimento moral sem considerar o elemento da crença, nem simular artificialmente a crença para produzir conhecimento moral. Para examinar essas hipóteses, a tese utiliza dilemas morais, envolvendo situações de tomada de decisão complexa, que são modelados através de equações de lógica modal para avaliar a capacidade da IA de resolver esses dilemas. A análise mostra que, embora a IA possa simular respostas a dilemas baseando-se em premissas lógicas, sempre há uma lacuna significativa devido à ausência de motivação moral e compreensão contextual, cruciais para decisões genuinamente morais. Incorporando o pensamento de Stuart Mill e Jesse Prinz, a pesquisa aprofunda a compreensão de como conceitos filosóficos tradicionais podem ser aplicados a problemas modernos, ilustrando que a empatia, a motivação moral humana e o julgamento ético são aspectos que a IA atualmente não pode replicar ou incorporar de forma autônoma. A conclusão reforça que a IA, por mais avançada que seja, não pode substituir o julgamento moral humano. As contribuições científicas desta pesquisa são substanciais, promovendo o debate sobre ética da IA esclarecendo os limites da automatização de decisões morais.

Palavras-chave: conhecimento moral; sistemas inteligentes; lógica.

Abstract

AZAMBUJA, Flávia Braga de. A study on the possibility of representing moral knowledge through logic for application in intelligent systems: a non-monotonic logic approach. Supervisor: Juliano Santos do Carmo. 2024. 116 pp. Thesis (PhD in Philosophy) – Institute of Philosophy, Sociology and Politics, Federal University of Pelotas.

This thesis investigates the capability of artificial intelligence (AI) systems to engage in morally significant decision-making. The general hypothesis posits that although AI is becoming ubiquitous in contexts requiring moral judgments, we cannot attribute moral responsibility to autonomous systems because moral knowledge cannot be fully represented or simulated by logic. The specific hypotheses explore various nuances of this central issue. It is argued that autonomous systems lack "moral motivation" and that, despite AI's ability to automate some human thought processes such as creativity and decision-making, it cannot generate authentic moral knowledge due to the impact of data bias and the conditional and contextual nature of knowledge. The thesis also maintains that it is not possible to establish a basis for moral knowledge without considering the element of belief, nor to artificially simulate belief to produce moral knowledge. To examine these hypotheses, the thesis utilises moral dilemmas involving complex decision-making situations, modelled through modal logic equations to assess AI's capability to resolve these dilemmas. The analysis shows that while AI can simulate responses to dilemmas based on logical premises, there is always a significant gap due to the absence of moral motivation and contextual understanding, crucial for genuinely moral decisions. Incorporating the thoughts of Stuart Mill and Jesse Prinz, the research deepens understanding of how traditional philosophical concepts can be applied to modern problems, illustrating that empathy, human moral motivation, and ethical judgement are aspects that current AI cannot replicate or autonomously incorporate. The conclusion reinforces that AI, no matter how advanced, cannot replace human moral judgement. The scientific contributions of this research are substantial, promoting the debate on AI ethics by clarifying the limits of automating moral decisions.

Keywords: moral knowledge; intelligent systems; logic.

Sumário

Capítulo 1: Introdução

Este capítulo apresenta o escopo do estudo, estabelecendo os objetivos gerais e específicos, bem como as hipóteses que direcionaram a investigação sobre a geração de conhecimento moral[1] através da lógica para aplicação em sistemas inteligentes. A escolha deste tema foi motivada pela crescente integração da inteligência artificial (IA) em contextos que exigem considerações morais, levantando questões fundamentais sobre a capacidade de máquinas em desempenhar funções que tradicionalmente requerem discernimento moral humano. As seções seguintes expõem as hipóteses gerais e específicas que guiam o estudo, articulando as nuances e os desafios associados à moralidade[2] em sistemas inteligentes. A justificação para esta pesquisa é detalhada, enfatizando a importância e a urgência de abordar a interseção entre ética e tecnologia avançada, num momento em que as capacidades da IA continuam a expandir-se rapidamente.

A presente tese tem como objetivo examinar as implicações de gerar conhecimento moral através da lógica, visando sua aplicação em sistemas inteligentes. O tema foi escolhido levando em consideração o pressuposto de que a moralidade é uma das discussões intelectuais mais desafiadoras para a humanidade. Com o avanço tecnológico, uma das questões amplamente debatidas é se devemos ou podemos ensinar moralidade às máquinas. O avanço dos sistemas de inteligência artificial, que permitem uma autonomia crescente em diversos aplicativos, torna essa questão cada vez mais relevante.

[1] Neste trabalho não irei investigar o conceito de conhecimento moral, mas tomarei como pressuposto que é possível conhecer verdades morais de maneira objetiva e justificável.

[2] O conceito de moralidade adotado nesta tese baseia-se nas perspectivas de Stuart Mill e Jesse Prinz, visando investigar a possibilidade de moralidade na inteligência artificial. O utilitarismo de Mill foca nas consequências das ações para determinar sua moralidade (MILL, J. S. Utilitarianism. London: Parker, Son, and Bourn, 1863), fornecendo um framework claro para avaliar as ações de sistemas de IA. Prinz argumenta que as emoções e contextos culturais são centrais para nossos julgamentos morais (PRINZ, J. The Emotional Construction of Morals. Oxford: Oxford University Press, 2007), o que é crucial para entender como IA pode ser treinada ou programada para replicar ou respeitar sensibilidades morais humanas.

Como Hipótese Geral (HG) desta tese, adotamos a seguinte premissa: à medida que a IA se torna onipresente, ela estará cada vez mais envolvida em situações novas e moralmente significativas. No entanto, não podemos atribuir responsabilidade moral aos sistemas inteligentes, pois o conhecimento moral não pode ser representado unicamente pela lógica.

1.1 Caracterização do Problema de Tese

Nesta seção, apresentamos a caracterização do problema de tese para o qual serão estabelecidas as diretrizes teóricas deste trabalho.

1.2 Hipótese Geral

À medida que a IA se torna onipresente, ela estará cada vez mais envolvida em situações novas e moralmente significativas. No entanto, não podemos atribuir responsabilidade moral a sistemas inteligentes, pois o conhecimento moral não pode ser representado pela lógica.

Hipóteses Específicas

São hipóteses específicas desta tese:

- Não é possível gerar "motivação moral[3]" em sistemas inteligentes.

- A inteligência artificial torna possível automatizar os processos de pensamento humano, criatividade e tomada de decisões.

- O viés dos dados pode ser gerado por sistemas inteligentes, resultando em conclusões errôneas e tendenciosas.

- Não é possível estabelecer uma base para o conhecimento sem considerar o elemento da crença, nem simular artificialmente a crença[4] para a produção do

[3] De acordo com Vasiliou (2016), o conceito de motivação moral refere-se ao estudo de como, por que e se os julgamentos morais – isto é, julgamentos de que alguma ação é correta, moral, ética ou virtuosa – motivam os agentes a agir. Este conceito explora a relação entre as crenças morais de um indivíduo e suas disposições comportamentais, investigando se os julgamentos morais inerentemente incentivam ações correspondentes e quais mecanismos psicológicos e filosóficos estão envolvidos nesse processo de motivação.

[4] "uma atitude mental de aceitação ou assentimento de que uma determinada proposição é verdadeira.

conhecimento moral.

- Não é possível usar ferramentas de ciências exatas, como a lógica e a matemática, por meio da filosofia científica, para sistematizar e gerar conhecimento moral em sistemas inteligentes.

- O conhecimento é condicional; saber quando aplicar um procedimento é tão importante quanto conhecer o procedimento (condicional).

- O conhecimento está relacionado ao contexto.

1.3 Objetivos

Esta seção detalha os objetivos que orientam o desenvolvimento desta tese, divididos em objetivo geral e objetivos específicos. O objetivo geral serve como o norte conceitual da investigação, enquanto os objetivos específicos delineiam as etapas concretas através das quais o estudo buscará alcançar sua meta principal. Juntos, estes objetivos estruturam o arcabouço analítico e metodológico do trabalho, definindo claramente o que a pesquisa pretende explorar, analisar e demonstrar em relação à aplicação de conhecimento moral em sistemas inteligentes.

Objetivo principal

Argumentar sobre a inviabilidade de gerar conhecimento moral através da lógica para aplicação em sistemas inteligentes.

Objetivos secundários

- Delimitar as possibilidades e limitações para a geração de conhecimento moral através da lógica;
- Analisar a importância das crenças na formação do conhecimento moral;
- Identificar os conceitos que determinam a impossibilidade de transmissão do pensamento e compreensão da moral em sistemas inteligentes;

É um estado psicológico que representa a forma como o mundo é percebido por um indivíduo." (AUDI, 2010)

- Investigar os processos de geração conhecimento moral em sistemas inteligentes no desenvolvimento de aprendizado de máquina, compreendendo o processo de representação da linguagem e do pensamento.
- Investigar as possibilidades de representação do conhecimento moral ou da moralidade.

1.4 Justificativa

O estudo da inviabilidade de produção de conhecimento moral através de algoritmos lógicos se justifica diante às seguintes questões:

- Podemos construir agentes morais [5]artificiais usando aprendizado de máquina?
- Existe a possibilidade de treinar a IA para identificar o bem e o mal, o certo e o errado e depois usá-la para nos ensinar moralidade?

A discussão sobre moralidade e ética na IA emerge como um dos dilemas mais complexos e intrigantes no cruzamento entre tecnologia e filosofia. Este desafio se acentua em um contexto dinâmico de avanços tecnológicos, onde os limites do aprendizado e da autonomia das máquinas são continuamente testados. Portanto, nosso debate central gira em torno da discussão sobre a possibilidade de incorporar moralidade em sistemas inteligentes emergentes, destacando-se como um campo de investigação fundamental. Esta tese aborda tal debate, reconhecendo a moralidade como sendo um dos temas intelectuais mais desafiadores para a humanidade.

A rápida evolução dos sistemas de IA, que oferecem crescente autonomia em variadas aplicações, sublinha a importância de se discutir como esses sistemas podem aderir, interpretar ou desenvolver princípios morais. No nosso entendimento, existe uma lacuna significativa na existência de padrões universais ou diretrizes específicas que possam ser utilizados para a implementação de normas humanas e valores morais em sistemas inteligentes. Portanto, o objetivo deste trabalho é explorar

[5] Entendemos como agente moral como sendo um ser que pode ser considerado responsável por suas ações, pois possui a capacidade de discernir o certo do errado, deliberar sobre diferentes opções, e agir de acordo com princípios éticos. Essa capacidade inclui a habilidade de entender e aplicar normas morais, refletir sobre as consequências de suas ações e tomar decisões com base em considerações éticas.

essa lacuna, propondo uma análise detalhada sobre as interações possíveis entre moralidade e IA, além dos desafios e implicações dessa integração.

Iniciamos a discussão sobre moralidade com base na abordagem de Tomasello (2015), que sugere a emergência da moralidade, uma característica da cooperação humana, em duas formas analógicas na natureza. Primeiramente, um indivíduo pode escolher agir em detrimento próprio para beneficiar outro, guiado por motivações como empatia, preocupação e generosidade. Alternativamente, indivíduos interagindo podem visar um benefício recíproco fundamentado em critérios objetivos, como justiça e equidade. Ambos os enfoques encontram respaldo no imperativo categórico kantiano, que orienta agir conforme uma máxima que poderia ser adotada como lei universal. Este princípio se aplica tanto à noção de auto sacrifício por razões altruístas quanto à formação de relações pautadas em justiça e equidade. No contexto do auto sacrifício, se a ideia de se doar pelo bem comum fosse universalizável, tal atitude seria congruente com a obrigação moral. Em relação à equidade e justiça, estes conceitos são inerentemente aplicáveis de forma universal, favorecendo interações que respeitam a autonomia e dignidade individual. Entretanto, a questão primordial reside no fato de que, em contextos que exigem justiça, frequentemente se observa uma dinâmica complexa entre as tendências cooperativas e competitivas entre diversos agentes. O esforço em prol da justiça implica na busca de um equilíbrio entre essas tendências, existindo múltiplos mecanismos possíveis para alcançá-lo, fundamentados em critérios distintos.

No contexto do desenvolvimento de sistemas inteligentes aptos a tomar decisões e interagir, é crucial que esses sejam concebidos para compreender, assimilar e incorporar as normas e valores morais da sociedade. Assim, o desafio central reside em alinhar a IA com os valores e normas morais humanos, considerando seu papel decisório e as consequências morais envolvidas. Tendo isso em vista, abordamos também o pensamento de John Stuart Mill, para avaliar a possibilidade de representar a moralidade em sistemas inteligentes, justificado por sua abordagem que enfatiza a complexidade e a contextualidade da formação do caráter humano. Mill argumenta que é impossível obter proposições realmente precisas sobre a formação do caráter apenas pela observação e experimentação de fenômenos complexos; é necessário decompor esses fenômenos em componentes simples e compreender as leis gerais que os regem. Além disso, Mill é conhecido por

sua abordagem utilitarista, essa perspectiva utilitarista é relevante ao avaliar a integração de valores morais em sistemas de IA, pois enfatiza a importância das consequências das ações e a necessidade de ponderar os impactos morais das decisões tomadas pelos sistemas inteligentes. A abordagem utilitarista de Mill complementa sua análise da formação do caráter, pois reforça a necessidade de compreender as leis gerais que governam os fenômenos morais e aplicar esses princípios de forma que promovam o bem-estar coletivo.

A IA deve ser capaz de decompor fenômenos morais complexos em componentes simples e compreender as leis gerais que os regem, ao mesmo tempo em que busca maximizar os resultados positivos e minimizar os negativos. Esses fatores tornam evidente a dificuldade de replicar a moralidade humana de maneira precisa e eficaz em sistemas de IA, dado que a moralidade envolve uma profunda compreensão das nuances e contextos humanos que vão além da capacidade atual da tecnologia.

No contexto da IA, isso significa que, para integrar valores morais, seria necessário primeiro entender os componentes básicos da moralidade humana e suas interações. Contudo, a formação do caráter humano envolve fatores não empíricos, como experiências subjetivas, contextos culturais e influências sociais, que são difíceis de quantificar e modelar. Essas nuances e a necessidade de adaptação constante a circunstâncias específicas tornam a tarefa de replicar a moralidade humana em sistemas inteligentes extremamente complexa e, potencialmente, inviável. Portanto optamos também por abordar o pensamento do filósofo Jesse Prinz, que enfatiza a importância das emoções na tomada de decisão humana, o que adiciona outra camada de complexidade. Segundo Jesse Prinz, as emoções influenciam profundamente nossos julgamentos morais, e entendemos que a capacidade da IA de replicar essas emoções e entender seu impacto nas decisões morais ainda é limitada. Portanto, integrar valores morais em IA requer uma compreensão profunda não apenas das interações morais básicas, mas também da influência das emoções e das complexas dinâmicas contextuais e culturais.

Outra consideração relevante é o avanço dos sistemas digitais e a habilidade de "aprendizado" dos sistemas baseados em IA, que devem ser sustentados por princípios éticos e morais. Contudo, ao admitir que estas tecnologias são moldadas

por especificações humanas, reconhecemos igualmente que podem manifestar vieses inerentes à condição humana, não agindo como entidades completamente autônomas.

Dada a compreensão de que essas tecnologias refletem vieses humanos, conforme discutido anteriormente, e reconhecendo que elas são desenvolvidas a partir de conjuntos de dados de treinamento com forte influência humana — sendo estes dados a base dos sistemas inteligentes —, torna-se imperativo que a comunidade acadêmica intensifique a pesquisa em ética aplicada ao desenvolvimento tecnológico. Isso se justifica pelo fato de que tais sistemas já podem estar realizando decisões de significativa relevância moral. Além disso, precisamos considerar também que por trás de uma decisão ética sempre existe a motivação moral humana, que tem papel decisivo na formação de julgamentos. Enquanto a distinção entre o certo e o errado, fundamentada em princípios morais, orienta inúmeras decisões, a presença de motivações morais nessas escolhas é inegável e necessita de atenção. Logo, entendemos que é necessária uma reflexão sobre a abordagem filosófica de ética e moralidade já que essas estão diretamente relacionadas à natureza humana.

Optamos por utilizar nesta pesquisa a lógica não monotônica, ao invés da lógica monotônica, para discutir a representação do conhecimento moral em sistemas inteligentes devido à necessidade de modelar a capacidade humana de revisar e atualizar crenças diante de novas informações. A lógica monotônica, que não permite a retração de conclusões previamente deduzidas, é insuficiente para capturar a dinâmica e a flexibilidade do raciocínio moral humano. O raciocínio moral frequentemente requer a reconsideração de julgamentos à luz de novos contextos e evidências. Em contraste, a lógica não monotônica reflete melhor essa adaptabilidade, permitindo que sistemas inteligentes ajustem suas inferências morais conforme novas informações e circunstâncias se apresentam, alinhando-se mais estreitamente com o funcionamento cognitivo humano..

1.5 Metodologia

Com base no contexto apresentado, foi utilizado o método do experimento mental ou experimentos do pensamento, por meio de modelos mentais e raciocínio não monotônico, para a argumentação da impossibilidade de representação de

conhecimento moral e motivação moral em sistemas inteligentes. Segundo Pereira (2015), esse método consiste em empregar situações imaginárias para auxiliar na compreensão ou previsão de como as coisas podem se comportar na realidade. Assim como os experimentos concretos, os experimentos mentais são ferramentas essenciais na construção do conhecimento científico. Essas situações imaginárias desempenham um papel importante na argumentação científica.

Dessa forma, foi possível identificar padrões de interpretação, por meio de categorias de definições apresentadas por pontos de vista significativos, através de um experimento mental. Foram testadas as possibilidades de representar e processar conceitos morais, motivação moral e emoções, utilizando a lógica modal, que permitiu argumentar em que condições seria possível que sistemas inteligentes incorporem conceitos abstratos de motivação moral, conhecimento moral e emoções a fim de que eles possam realizar análises éticas mais sofisticadas e tomar decisões alinhadas com princípios morais e emocionais humanos.

Nos experimentos, foram considerados cenários hipotéticos envolvendo dilemas morais que abordam princípios morais universais, que se aplicam em todos os mundos possíveis, independentemente das circunstâncias específicas de cada mundo considerando questões éticas complexas.

Ao realizar o experimento mental, foram avaliadas as seguintes questões:

- Representação de Conceitos Morais: se é possível representar e compreender os princípios morais relevantes incorporando conceitos como beneficência, não maleficência, autonomia e justiça, utilizando a lógica modal para expressar esses princípios de forma coerente.
- Incorporação de motivação moral e emoções: se é possível representar e processar motivação moral e emoções relevantes para a tomada de decisões éticas tais como: compaixão, empatia e preocupação.

Procedimentos adotados:

- Construção do conjunto inicial de regras morais: foi definido um conjunto inicial de regras morais básicas, formuladas de forma a capturar princípios éticos fundamentais, como evitar causar danos,

promover o bem-estar geral, respeitar direitos individuais, entre outros.

- Modelagem dos dilemas morais: foram desenvolvidos 3 dilemas morais complexos envolvendo conflitos de princípios éticos. Cada dilema foi cuidadosamente elaborado para desafiar o conhecimento moral e a motivação moral a serem representados pela lógica não-monotônica.
- Representação do conhecimento moral: foi utilizada a lógica não-monotônica para representar o conhecimento moral inicial, codificando as regras morais básicas.
- Análise dos dilemas morais: foram analisados os dilemas morais codificados em um sistema de lógica não-monotônica, considerando as informações disponíveis e as regras morais iniciais. Também foram analisados com base nos conceitos filosóficos discutidos nesta tese.

Capítulo 2: Abordagem filosófica sobre ética e moralidade focada no desenvolvimento da inteligência artificial.

Este capítulo procura sintetizar teorias filosóficas de John Stuart Mill e Jesse Prinz, investigando como essas teorias podem ser transpostas e adaptadas para enfrentar os dilemas éticos emergentes na era digital. Abordaremos conceitos de moralidade de acordo com esses filósofos, incluindo uma exploração do conceito de motivação moral, analisando como e por que as pessoas decidem agir de maneiras que consideram corretas ou erradas. Discutiremos as potencialidades e as limitações de aplicar os pensamentos de Mill e Prinz ao desenvolvimento de tecnologias de IA. Isso inclui um diálogo propositivo entre filosofia e tecnologia, visando uma compreensão mais aprofundada de como princípios éticos tradicionais podem ser relevantes no contexto tecnológico contemporâneo.

2.1 A Moralidade segundo John Stuart Mill

Dada a importância do tema moralidade dentro do contexto da representação do conhecimento moral, optamos por abordar a questão de ética e moralidade, inicialmente, através da análise do pensamento de John Stuart Mill, em "A Lógica das Ciências Morais", que nos oferece uma perspectiva relevante para o debate e que pode contribuir para entender como a formação do caráter humano pode interferir na IA. Em sua obra, Stuart Mill (2020), enfatiza a importância de desmembrar fenômenos complexos em componentes mais simples para compreendê-los analiticamente, uma abordagem que entendemos pode ser paralelamente aplicada ao desafio de integrar valores morais em sistemas inteligentes (MILL, 2020).

Assim como Stuart Mill sugere que a compreensão das leis gerais da mente é fundamental para deduzir os efeitos de circunstâncias específicas na formação do caráter, argumenta-se que um conhecimento aprofundado dos princípios éticos é crucial ao considerar se a inteligência artificial pode incorporar e refletir valores morais. Entretanto, isso implicaria em determinar a viabilidade e os métodos pelos quais valores morais podem ser efetivamente integrados e manifestados em sistemas inteligentes. Neste contexto, consideramos que a abordagem para compreender a formação do caráter humano oferece uma perspectiva valiosa para o aprofundamento em ética aplicada ao desenvolvimento tecnológico. John Stuart Mill, afirma:

> Visto, portanto, que é impossível obter apenas pela observação e pelo experimento proposições realmente precisas a respeito da formação do caráter, somos necessariamente levados àquele modo de investigação que, mesmo que não fosse indispensável, teria sido o mais perfeito e cuja extensão é um dos principais objetivos da filosofia; isto é, àquele modo que tenta seus experimentos não sobre os fatos complexos, mas sobre os fatos simples que os compõem e que, após estabelecer as leis das causas cuja composição dá origem aos fenômenos complexos, considera se estas não explicam e dão conta das generalizações aproximadas que foram formadas empiricamente a respeito das sequências desses fenômenos complexos. As leis da formação do caráter são, em suma, leis derivadas resultantes das leis gerais da mente e, para obtê-las, devemos deduzi-las dessas leis gerais, supondo um conjunto dado um conjunto qualquer de circunstâncias e considerando então qual será, de acordo com as Leis da Mente, a influência dessas circunstâncias na formação do caráter (Mill, 2020, p-78-79).

Adicionalmente, se considerarmos a ênfase de Stuart Mill na interação entre leis mentais gerais e circunstâncias específicas, ressaltamos a importância de considerar o contexto no qual a IA opera. Isso implica que a incorporação de valores morais em sistemas inteligentes precisaria levar em conta todas as nuances e variações das situações reais, assim como a interação desses sistemas com humanos e outras entidades, reforçando a ideia de que o desenvolvimento de IA ética requer um esforço conjunto entre a filosofia, a tecnologia e a ciência. Esta abordagem visa uma compreensão profunda tanto das leis gerais que regem a moralidade quanto das especificidades contextuais que influenciam a tomada de decisões morais. Neste contexto, a reflexão de Stuart Mill sobre o conhecimento das leis da formação do caráter e sua aplicabilidade prática se torna particularmente relevante. Ele observa:

> É certo que, por mais completa que seja a determinação das leis da formação do caráter, seria infundado esperar poder conhecer tão exatamente as circunstâncias de um caso dado qualquer para sermos capazes de predizer positivamente o caráter que seria produzido no caso. Mas devemos recordar que um grau de conhecimento insuficiente para autorizar uma predição efetiva é, frequentemente, de grande valor prático. Pode-se dispor de um grande poder de influenciar os fenômenos por meio de um conhecimento bastante imperfeito das causas pelas quais eles são, em qualquer instância dada, determinados. É suficiente saber que certos meios têm uma tendência para produzir um dado efeito e que outros têm a tendência para frustrá-lo. Quando as circunstâncias de um indivíduo ou de uma nação estão, em um grau considerável, sob nosso controle, podemos, pelo conhecimento das tendências, estar capacitados a arranjar essas circunstâncias de uma maneira muito mais favorável aos fins que desejamos do que o arranjo que elas assumiriam por si mesmas. Esse é o limite de nosso poder, mas, dentro desse limite, é um poder dos mais importantes (Mill, 2020, p-136).

A observação de Stuart Mill sobre a complexidade da previsão acurada do impacto de circunstâncias específicas na formação do caráter humano mostra um desafio a ser enfrentado pelas tecnologias de IA em replicar sem viés a complexidade moral humana. Esta citação evidencia que pode existir uma limitação em simular dinâmicas humanas complexas, pois é necessário considerar a dificuldade e os obstáculos em reproduzir, de forma imparcial, os princípios morais humanos.

Entretanto Stuart Mill, em seus trabalhos sobre a lógica e a filosofia da ciência também considera a complexidade de aplicar métodos científicos rigorosos às ciências morais, pois embora acreditasse na possibilidade de aplicar a lógica e o rigor científico ao estudo desses campos, reconhecia os desafios únicos que eles apresentam em comparação com as ciências físicas. Para Stuart Mill, o processo de compreensão das complexidades do comportamento humano e da organização social inicia com a observação atenta de ações individuais, instituições e fenômenos morais, buscando identificar padrões e princípios que regem tais comportamentos e estruturas. Esta abordagem, segundo ele, não apenas facilita a generalização a partir de casos específicos, mas também permite a formulação de leis mais abrangentes que podem ser aplicadas a um espectro mais amplo de situações.

> [...] é evidente que a Sociologia, considerada como um sistema de deduções a priori, não pode ser uma ciência de predições positivas, mas apenas de tendências. Podemos ser capazes de concluir, a partir das leis da natureza humana aplicadas às circunstâncias de um certo estado da sociedade, que uma causa particular irá operar de uma certa maneira a menos que seja contrariada. Entretanto, jamais podemos estar seguros em relação à extensão ou ao grau em que irá assim operar ou afirmar com certeza que nunca será contrariada, pois raramente conhecemos, mesmo aproximadamente, todos os agentes que podem coexistir com ela e ainda menos podemos calcular o resultado coletivo de tantos elementos combinados (Mill, 2020, P-123).

Na teoria de Stuart Mill todo conhecimento deriva da experiência em que observações específicas são usadas para formular generalizações mais amplas. Entender a mente humana e as sociedades exige uma análise cuidadosa das experiências e dos fenômenos observáveis, a partir dos quais as leis gerais poderiam ser deduzidas. Além disso, a compreensão das relações causais e das condições

específicas que afetam os fenômenos sociais e morais é muito importante pois, identificar as causas e os efeitos das ações e decisões humanas, bem como as circunstâncias que influenciam tais eventos pode contribuir para uma análise mais precisa das dinâmicas sociais e comportamentais. Na psicologia e nas ciências sociais, isso significaria coletar dados sobre comportamentos, emoções e interações sociais para identificar padrões e princípios subjacentes.

Outra importante contribuição de Stuart Mill, trata da necessidade de flexibilidade nas ciências morais, argumentando que as leis gerais devem ser adaptáveis a casos particulares, reconhecendo a diversidade de contextos e circunstâncias que podem influenciar a manifestação dessas leis. Essa adaptabilidade é crucial para a aplicação prática dos conhecimentos adquiridos, permitindo uma resposta mais eficaz às variadas situações humanas e sociais.

Para Stuart Mill, as ciências morais não se limitam a uma função teórica, elas têm um papel prático significativo na melhoria da sociedade e entender as leis que regem o comportamento humano e a organização social é, para ele, fundamental para o aprimoramento da condição humana e da sociedade como um todo. Apesar de seus desafios únicos as ciências morais podem ser abordadas de forma lógica e científica, através de uma metodologia empírica, analítica e causal, com o objetivo de compreender profundamente o comportamento humano e a sociedade, e de utilizar esse conhecimento para promover o bem-estar coletivo. Essa abordagem não só destaca a importância de uma fundamentação empírica na análise das questões morais e sociais, mas também sublinha o potencial dessas ciências para contribuir para melhorias tangíveis na vida social.

Entretanto, a abordagem de Stuart Mill à moralidade e à ética se mostrou fortemente influenciada por sua adesão ao utilitarismo. O utilitarismo de Mill é baseado no conceito da maior felicidade, ou seja, a ação moralmente correta seria aquela que proporciona maior felicidade ao maior número de pessoas possível, baseado na premissa de que as ações são corretas na medida em que tendem a promover a felicidade e erradas na medida em que tendem a produzir o oposto da felicidade. Essa posição também foi fundamentada em princípios empíricos e consequencialistas, avaliando a moralidade das ações pelos seus resultados.

A abordagem de Stuart Mill ao utilitarismo, com foco na maximização da felicidade geral, apresenta um desafio intrigante quando consideramos sua aplicação através da IA para a tomada de decisão moral. Embora a IA possa processar e analisar uma vasta quantidade de dados rapidamente, a complexidade de prever todas as possíveis consequências de uma ação e avaliar seu impacto global na felicidade é um grande desafio porque isso envolve não apenas entender as consequências imediatas, mas também as repercussões a longo prazo, que podem ser extremamente difíceis de modelar. Sendo a ideia de felicidade subjetiva e considerando que essa ideia varia entre indivíduos e culturas, determinar um padrão uniforme de felicidade que possa ser aplicado universalmente é problemático, pois a IA teria que se basear em alguma forma de medição da felicidade, o que poderia incluir indicadores quantitativos e qualitativos. Ou seja, a complexidade reside em como programar esses padrões na IA de forma que ela possa avaliar adequadamente a felicidade em diferentes contextos. A abordagem de Stuart Mill baseia-se em avaliar as ações pelos seus resultados. Enquanto a IA pode ser extremamente eficaz em identificar padrões e prever resultados com base em dados históricos, o desafio surge na aplicação de princípios empíricos a situações novas e sem precedentes. Além disso, a avaliação moral não se limita a uma análise consequencialista; ela também envolve considerações deontológicas e éticas que podem não ser facilmente quantificáveis ou previsíveis por algoritmos.

Na prática, em muitas situações, o que maximiza a felicidade para a maioria pode prejudicar uma minoria, levantando questões éticas sobre justiça e direitos individuais. Programar uma IA para navegar nesses dilemas morais, onde interesses conflitantes devem ser equilibrados, é uma tarefa complexa. A decisão "correta" nem sempre é clara, e a IA pode ter dificuldade em fazer julgamentos que exigem empatia e compreensão dos valores humanos.

A ligação entre a abordagem de Stuart Mill ao utilitarismo e sua contribuição à lógica, no contexto da avaliação de situações morais pela IA, ressalta um aspecto crucial da complexidade em aplicar teorias éticas tradicionais ao desenvolvimento e funcionamento da IA. Stuart Mill, em sua obra, não propôs um modelo lógico formal específico para a representação moral, mas sim enfatizou a importância do empirismo, do utilitarismo e da investigação científica para entender e avaliar o comportamento humano e a sociedade. Isso sugere uma abordagem à ética que é mais flexível e

adaptável do que as formulações rígidas oferecidas pela lógica formal.

No contexto da IA e do utilitarismo de Stuart Mill, essa perspectiva empirista e consequencialista sobre a moralidade e a ética é particularmente relevante. A IA, por sua natureza, opera por meio da análise de dados, aprendizado a partir de exemplos específicos (empirismo) e tentativas de maximizar ou otimizar certos resultados (consequencialismo). Quando consideramos a capacidade da IA de avaliar todas as possibilidades envolvidas na tomada de decisão moral sob a luz do utilitarismo, o foco de Mill na indução e na observação de casos particulares para chegar a conclusões gerais sobre o bem-estar humano se torna especialmente pertinente.

A abordagem de Stuart Mill implica que a tomada de decisão moral e ética pode ser fundamentada na observação e análise de casos particulares, avaliando suas consequências para o bem-estar humano. Isso ressoa com a maneira como uma IA poderia teoricamente operar ao aplicar princípios utilitaristas: analisando dados sobre situações específicas, prevendo as consequências de várias ações e escolhendo aquelas que promovem a maior felicidade geral. No entanto, isso também destaca as limitações e desafios mencionados anteriormente, como a dificuldade de definir e medir a "felicidade", a complexidade de prever todas as possíveis consequências a longo prazo e a necessidade de equilibrar interesses conflitantes.

Assim, ao contextualizar a contribuição de Stuart Mill à lógica dentro do quadro da aplicação de IA à tomada de decisão ética, percebemos que, embora Stuart Mill não tenha desenvolvido um modelo lógico formal para a ética, sua ênfase na indução, empirismo e consequencialismo oferece uma base teórica que, em certo sentido, é compatível com os métodos operacionais da IA. Isso reforça a ideia de que a aplicação de teorias éticas na IA requer uma abordagem multidimensional, que considere tanto os princípios filosóficos quanto as capacidades e limitações tecnológicas.

2.2 A Visão de Jesse J. Prinz sobre Moralidade

Considerando o avanço das pesquisas em inteligência artificial, avaliamos a relevância de abordar a temática da moralidade através de perspectivas filosóficas atuais. Neste contexto, destaca-se a análise da concepção de moralidade proposta pelo filósofo contemporâneo Jesse J. Prinz (2007) que em seu ensaio "Is Morality Innate?" investiga a questão crítica de se a moralidade é uma característica inata ao

ser humano, um tema de significativa importância para os estudos que buscam compreender até que ponto a inteligência artificial pode simular ou replicar o discernimento moral humano.

Jesse Prinz defende uma posição bastante distinta sobre a origem da moralidade, contrariando a ideia de que a moralidade é inata, Prinz argumenta que ela é, na verdade, culturalmente construída. Segundo Prinz, não existem "módulos morais" inatos no cérebro humano, em vez disso, nossos juízos morais são o resultado de nossas experiências e do ambiente cultural em que estamos inseridos.

Embora a busca por uma explicação biológica para a moralidade possa elucidar a universalidade das normas morais, também suscita preocupações éticas, especialmente considerando o histórico da eugenia. Apesar disso, a inclinação a atribuir à humanidade uma faculdade moral inata continua sendo uma ideia atraente, sugerindo uma visão otimista sobre a natureza humana. Jesse Prinz desafia a visão de uma moralidade inata ao destacar os perigos de adotar explicações biológicas sem uma análise crítica. Embora a ideia de uma capacidade moral inata possa ser sedutora por sugerir uma característica única da espécie humana e explicar a universalidade das normas morais, Prinz argumenta que não há evidências convincentes para sustentá-la. Ele propõe que a moralidade surge de capacidades evolutivas desenvolvidas para outros fins, não sendo derivada de uma faculdade moral específica, mas sim de nossa complexa arquitetura cognitiva. Essa perspectiva reconhece a variedade e adaptabilidade dos sistemas morais, promovendo uma abordagem mais diversificada no estudo da moralidade, que considera tanto nossas inclinações naturais quanto nossa capacidade de crescimento e aprendizado moral. Prinz enriquece o debate ao sugerir que a busca pelo bem vai além da existência de capacidades inatas, incorporando nossa capacidade de aprendizado, adaptação e cultivo de preocupações morais. Essa abordagem desafia visões simplistas e promove uma compreensão mais profunda e holística da natureza da moralidade.

> *Quero combater essa ideia sedutora. Não nego que a moralidade seja ecumênica, mas penso que ela não é inata — pelo menos que o estado atual das evidências não é convincente. A moralidade, como todas as capacidades humanas, depende de ter predisposições biológicas particulares, mas nenhuma dessas, eu argumento, merece ser chamada de faculdade moral. A moralidade é um subproduto — acidental ou inventado — de faculdades que*

> *evoluíram para outros propósitos. Como tal, a moralidade é consideravelmente mais variável do que o programa nativista poderia nos levar a pensar, e também mais versátil. É emocionante ver cientistas cognitivos demonstrando tanto interesse em moralidade, mas essa tendência carrega o risco de reificação. Quero argumentar por uma história mais confusa, mas espero que essa história leve a uma melhor compreensão de como passamos a nos importar com o bem. Penso que a história nativista superestima a decência humana e subestima o potencial humano (PRINZ, 2007, p. -1).*[6]

Assim, Jesse Prinz contesta a concepção de que a moralidade é inata, sugerindo, ao invés disso, que é um fenômeno intrincado resultante de várias interações e predisposições biológicas, mas não atribuível exclusivamente a uma faculdade moral inata específica. Um dos fundamentos centrais de sua argumentação é o emocionismo, a proposição de que as emoções desempenham um papel crucial na formação dos juízos morais. Conforme Prinz, as emoções não são meramente reações aos juízos morais: elas são parte integrante desses juízos. Isso implica que, ao avaliarmos uma ação como moral ou imoral, estamos, na realidade, expressando uma resposta emocional baseada em nossa aprendizagem e contexto cultural. Esse ponto de vista transfere a origem dos juízos morais de uma base racional e inata para uma que se desenvolve a partir das experiências emocionais e sociais do indivíduo.

Ao considerar a posição de Jesse Prinz sobre a moralidade como uma construção cultural e não um traço inato, pode-se estabelecer um interessante paralelo com os desafios enfrentados pela IA ao tentar representar ou simular a moralidade humana. Assim como John Stuart Mill apontou a dificuldade de prever ações humanas devido à interação de múltiplas causas primárias e secundárias, representar algoritmicamente a moralidade apresenta-se como um problema, pois, conceitualmente, de acordo com os autores, a moralidade humana não é determinada apenas por regras claras e decisões binárias, ela é matizada, dependente do contexto e muitas vezes contraditória.

---

[6] Texto original: “I want to combat this alluring idea. I do not deny that morality is ecumenical, but I think it is not innate—at least that the current state of evidence is unpersuasive. Morality, like all human capacities, depends on having particular biological predispositions, but none of these, I submit, deserves to be called a moral faculty. Morality is a byproduct—accidental or invented—of faculties that evolved for other purposes. As such, morality is considerably more variable than the nativism program might lead us to think, and also more versatile. It is exciting to see cognitive scientists taking such an interest in morality, but that trend carries the risk of reification. I want to argue for a messier story, but I hope that the story leads to better understanding of how we come to care about the good. I think the nativist story oversells human decency, and undersells human potential”. (PRINZ, 2007, p -1)

Além disso, Jesse Prinz aborda as propriedades morais sob a lente do internalismo motivacional, que sustenta que os juízos morais estão intrinsecamente ligados à motivação para agir de acordo com esses julgamentos. Ou seja, ao reconhecer uma ação como moralmente correta, um indivíduo se sente motivado a executar tal ação. Esta visão enfatiza o papel ativo das emoções e da cultura na formação de nossas concepções morais e na motivação para agir moralmente. A relação entre juízos morais e emoções é, para Prinz, uma evidência contra a noção de uma moralidade inata. Ele argumenta que se nossos julgamentos morais são tão profundamente influenciados por nossas respostas emocionais, que variam significativamente entre culturas e indivíduos, então a moralidade não pode ser considerada uma característica inata da espécie humana. Em vez disso, deve ser vista como um conjunto de práticas e crenças moldadas pela evolução cultural e pelo desenvolvimento individual.

O autor fornece três exemplos para respaldar sua tese de que as normas morais estão enraizadas em emoções morais. Primeiramente, ele menciona um estudo conduzido por Weatley e Haidt (2005) apud (Prinz, 2007), no qual os participantes foram hipnotizados para associar uma palavra neutra a uma emoção de repulsa. Como resultado, esses participantes passaram a emitir avaliações morais mais negativas em relação a indivíduos mencionados em histórias contendo essa palavra. Esse experimento demonstra como a manipulação das emoções pode influenciar diretamente os julgamentos morais. Em seguida, Jesse Prinz discute o déficit emocional observado em psicopatas, colocando que esses indivíduos apresentam uma profunda falta de emoções negativas, incluindo aquelas associadas a juízos morais enfrentando, consequentemente, dificuldades significativas em formar julgamentos morais e em distinguir entre regras morais e convencionais. Por fim, Jesse Prinz destaca uma correlação conceitual entre emoções e julgamentos morais. Ele ilustra isso com o exemplo de alguém que, embora acredite que uma ação maximizaria a felicidade, ainda pode considerá-la moralmente inadequada. Da mesma forma, alguém que reconhece uma ação como contraditória quando universalizada não necessariamente a considera moralmente má. Esses casos demonstram como as emoções influenciam diretamente as percepções sobre o caráter moral das ações. Esses exemplos fornecidos reforçam sua argumentação de que as emoções desempenham um papel crucial na formação de julgamentos morais, enfatizando a

importância de considerar o aspecto emocional na compreensão da moralidade humana.

Dentro deste contexto, indagamos se a Inteligência Artificial seria capaz de simular o aspecto da moralidade humana no qual o reconhecimento de uma ação como moralmente correta intrinsecamente motiva o indivíduo a executá-la. Considerando a relevância das emoções e da cultura na formação de nossas concepções morais e na motivação para agir de acordo com elas, questionamos se é viável desenvolver sistemas de IA capazes de simular esses processos emocionais complexos e responder a eles de maneira comparável ao comportamento humano. A investigação sobre como a IA poderia reproduzir o internalismo motivacional, que associa juízos morais à motivação para agir conforme esses julgamentos, nos conduz a uma reflexão mais ampla sobre o papel fundamental da motivação moral na ética, ao contribuir para a compreensão das razões subjacentes ao comportamento humano em situações morais.

Adicionalmente, se considerarmos a visão de Prinz de que a moralidade é uma construção cultural altamente dependente de respostas emocionais individuais, surge uma questão fundamental: pode uma entidade que não experimenta emoções de forma genuína participar autenticamente de práticas morais? Isso não apenas sublinha uma lacuna na capacidade algorítmica de simular a moralidade humana, mas também questiona a adequação de tais sistemas em contextos onde decisões morais precisam considerar compreensão empática e respostas emocionais profundas.

2.3 O problema da Motivação Moral

A discussão sobre motivação moral ocupa um lugar central na filosofia ética, explorando como e por que as pessoas decidem agir de maneiras que consideram corretas ou erradas. Esse tema ganha uma dimensão particularmente rica quando abordado a partir das perspectivas de John Stuart Mill e Jesse Prinz, cujas teorias oferecem contrastes iluminadores e complementares sobre o caráter e a moralidade. Stuart Mill, com sua ênfase no utilitarismo e na felicidade como a maior virtude, propõe que a motivação moral se baseia em promover o maior bem para o maior número. Por outro lado, Jesse Prinz argumenta a partir de uma perspectiva contemporânea que

desafia a noção de uma moralidade inata e sugere que nossas decisões morais são profundamente influenciadas por nosso ambiente cultural e emoções pessoais, enfatizando assim a relatividade cultural e a subjetividade das normas morais. Essas abordagens não apenas moldam nosso entendimento de como as normas morais são formadas e seguidas, mas também como elas são vivenciadas e justificadas pelos indivíduos em suas vidas diárias, destacando a importância de considerar tanto a universalidade das regras morais quanto a particularidade das experiências individuais.

Compreender a motivação moral humana pode ser útil para avaliar a possibilidade e a viabilidade de representar a motivação moral em sistemas inteligentes. Além disso, essa compreensão pode contribuir para a análise da natureza da responsabilidade moral e para a determinação da capacidade dos sistemas de serem moralmente responsáveis por suas ações.

O problema da motivação refere-se à relação entre o julgamento moral e a motivação para agir de acordo com esses julgamentos. Compreender esse problema é crucial tanto para a metaética quanto para a psicologia moral contemporâneas, bem como para a representação lógica do conhecimento em inteligência artificial. Uma questão central na investigação da motivação moral é determinar se as pessoas agem moralmente devido às suas próprias motivações internas ou se são influenciadas por fatores externos, como pressão social ou medo de punição.

Vários filósofos apresentaram perspectivas diferentes sobre a motivação moral. Para Kant (2011), a motivação moral genuína baseia-se no cumprimento do dever moral, guiado por princípios racionais e inerentes à própria vontade. Kant concorda com Hume (2009) quanto à importância da motivação para a moralidade, pois sem ela não haveria propriamente a ação com conteúdo moral. No entanto, a posição de Kant envolve uma sobreposição da sensibilidade com o entendimento, articulada no conceito do sentimento de respeito. A noção de princípio moral de Kant estabelece que o comportamento verdadeiramente moral é aquele que se baseia em algum tipo de julgamento normativo de dever, sendo consistente ou motivado por um princípio ou regra justificatória.

O pensamento de David Hume (2009) em relação à motivação moral enfatiza

significativamente a interação entre a moral e seu impacto direto nas paixões e comportamentos humanos. Para Hume, a moralidade transcende a esfera do intelecto para atuar como um agente transformador das ações e reações individuais. Ele propõe que a relevância da moral está em sua capacidade de influenciar diretamente nossas emoções e, consequentemente, nossas ações. A importância da moral é enfatizada por Hume no início do livro três onde coloca que “... A moral é um tema que nos interessa mais que qualquer outro” (Hume, 2009, L3, p-495). Hume questiona a ideia de que apenas a razão pode estabelecer padrões morais, e salienta a importância dos sentimentos e emoções na definição do que consideramos moralmente correto ou incorreto. Hume explora a questão de saber se distinguimos vício de virtude por meio de nossas ideias, ou seja, a razão, ou por nossas impressões, ou seja, os sentimentos. Ele afirma que a moral ativa as paixões e incentiva ou restringe comportamentos, enquanto a razão isoladamente mostra-se incapaz de exercer tal influência. Por meio de sua análise, Hume defende que entender as raízes morais e seu efeito prático sobre o comportamento humano é essencial para o estudo da filosofia moral. Portanto, ele argumenta que a filosofia moral deve considerar como a moralidade influencia práticas humanas através das paixões, sublinhando uma "necessidade motivacional" que liga as ideias de certo e errado à disposição para agir de acordo com esses julgamentos.

A motivação no pensamento de Stuart Mill (2020) aparece dentro do utilitarismo, uma teoria ética consequencialista que considera as consequências das ações como critério principal para determinar sua moralidade, onde fatores externos, como normas sociais e educação moral, podem influenciar a motivação moral ao moldar os valores e as motivações dos indivíduos. Kohlberg (2013), sustenta que a motivação moral é influenciada tanto por fatores internos quanto externos, incluindo a socialização, a interação com os outros e a exposição a diferentes perspectivas morais ao longo do tempo. Hume argumenta que os julgamentos morais não são derivados da razão, mas sim baseados em nossos sentimentos subjetivos de aprovação ou desaprovação. Ele afirmou que as distinções morais surgem de nossos sentimentos de aprovação moral (elogio) ou desaprovação (culpa) em relação a certas ações ou traços de caráter. Esses sentimentos não são baseados em nenhum princípio moral objetivo ou universal, mas sim nas respostas pessoais e subjetivas do indivíduo.

O papel da motivação moral na programação de sistemas inteligentes ainda é um tema complexo e em evolução. À medida que os sistemas inteligentes se tornam mais avançados e integrados em vários aspectos da sociedade, há um interesse crescente em garantir que esses sistemas possuam comportamento ético e capacidade de tomada de decisão e, de acordo com o que temos estudado, envolve o fator motivação.

Entre as diferentes abordagens sobre a possibilidade de moldar o comportamento desses sistemas e orientá-los para ações moralmente aceitáveis, Ryan e Stahl (2021) propõem algumas diretrizes deveriam ser consideradas, tais como: a importância das estruturas éticas, ou seja, designers e programadores de sistemas inteligentes estabeleceriam estruturas ou diretrizes éticas que delineariam as motivações morais desejadas para esses sistemas, o que envolveria a definição dos princípios e valores que devem orientar o comportamento e a tomada de decisão do sistema; a adoção de abordagens baseadas em regras, ou seja, inserir motivação moral em sistemas inteligentes através de programação baseada em regras, onde o sistema é programado com regras predefinidas ou princípios morais aos quais deve aderir, podendo ser baseados em teorias éticas estabelecidas ou normas sociais; a importância do aprendizado e adaptação, ou seja, projetar sistemas inteligentes para aprender motivações morais com base em feedback e em experiências retiradas da análise de dados, avaliando os resultados de suas ações e ajustando seu comportamento para se alinhar aos padrões morais desejados; alinhamento de valores, ou seja, garantir que as motivações morais dos sistemas inteligentes se alinhem com os valores humanos e as expectativas da sociedade considerando diversas perspectivas e engajando-se em discussões éticas durante as fases de desenvolvimento e programação; mecanismos de responsabilidade e supervisão, o que envolveria o estabelecimento de estruturas regulatórias, códigos de conduta ou processos de auditoria externa para monitorar e avaliar o comportamento do sistema, garantindo que permaneçam alinhados com os padrões morais e que quaisquer problemas ou violações éticas possam ser resolvidos. Entretanto, todas as abordagens mencionadas para incorporar motivação moral em sistemas inteligentes inevitavelmente dependem da interferência humana. Isso introduz um desafio significativo, visto que os vieses e perspectivas pessoais dos designers e programadores podem influenciar a programação moral dos sistemas. Tais vieses não

necessariamente refletem uma motivação moral universal ou os valores éticos compartilhados por diferentes culturas e sociedades. Este aspecto ressalta uma preocupação crítica: embora as diretrizes propostas por Ryan e Stahl (2021) busquem orientar o comportamento dos sistemas inteligentes de maneira ética, a subjetividade humana na definição dessas estruturas éticas pode comprometer a objetividade e a universalidade das motivações morais implantadas nos sistemas. Portanto, a tarefa de desenvolver sistemas inteligentes que não apenas compreendam, mas também ajam de acordo com princípios morais universalmente aceitos, exige um exame mais aprofundado das metodologias de design e programação, bem como uma reflexão contínua sobre a natureza da ética e moralidade que desejamos que essas tecnologias reflitam.

A representação lógica da motivação moral é uma área de pesquisa ativa no campo da ética computacional. Embora não haja uma abordagem única e amplamente aceita considerada o "estado da arte", vários pesquisadores propuseram estruturas e modelos formais para representar a motivação moral de maneira lógica ou computacional.

Alguns desses estudos incluem as pesquisas de Wallach et al. (2009), Floridi, (2011), Mittelstadt et al. (2016), Icard et al., (2017), Dignum (2018) e Jobin (2021). Esses pesquisadores investigam tópicos como lógica deôntica, mecanismos de raciocínio ético, modelos baseados em valores, web semântica, ontologias e também exploram a incorporação de dados empíricos, técnicas de aprendizado de máquina e processamento de linguagem natural para aprimorar a representação lógica da motivação moral em sistemas computacionais.

Capítulo 3: Modelos Mentais: Fundamentos Teóricos e Implicações

Este capítulo visa aprofundar a compreensão dos modelos mentais, uma noção fundamental na psicologia cognitiva que desempenha um papel essencial na maneira como os indivíduos percebem, interpretam e interagem com o mundo ao seu redor. A concepção inicial de modelos mentais foi proposta por Kenneth Craik (1967), que postulou que os indivíduos criam representações internas simplificadas de sistemas externos para facilitar o raciocínio e a previsão de eventos futuros. Esta abordagem foi expandida por Philip Johnson-Laird, cujos trabalhos desenvolveram uma estrutura teórica robusta que detalha a formação, manipulação e função dos modelos mentais na cognição humana.

Este capítulo delineará a natureza multifacetada dos modelos mentais, enfatizando como essas representações são construídas e utilizadas pelos indivíduos para simular e entender complexidades do mundo real. A discussão será ancorada na "Teoria dos Modelos Mentais", conforme elaborada por Johnson-Laird, e complementada por investigações subsequentes que exploram as interações entre conhecimento prévio, inferência e percepção.

Adicionalmente, será abordado o "viés da crença" dentro do contexto dos modelos mentais, um fenômeno pelo qual as predisposições existentes dos indivíduos podem distorcer a formação e a manipulação dessas representações cognitivas.

3.1 Modelos Mentais

A ideia inicial de modelos mentais remonta ao psicólogo escocês Kenneth Craik (1967), que sugeriu que a percepção constrói "modelos em pequena escala" da realidade que são usados para antecipar eventos e raciocinar. A teoria dos modelos mentais de Craik identifica três tipos de constructos representacionais: modelos mentais, imagens e proposições. Entre eles, os modelos mentais e as imagens são considerados representações de alto nível e essenciais para compreender a cognição humana. No entanto, esses modelos mentais são implementados por meio de um código proposicional. Nas representações proposicionais presentes nos modelos mentais, a linguagem desempenha um papel importante, pois captura os conceitos subjacentes a uma situação. Assim, o conteúdo ideacional da mente,

independentemente da modalidade original em que a informação foi encontrada, é representado por meio de representações linguísticas.

A teoria dos modelos mentais, proposta por Kenneth Craik em 1943, fundamenta-se na ideia de que o sistema nervoso humano funciona como uma máquina capaz de simular eventos externos da realidade, essencialmente modelando a realidade para produzir pensamentos e explicações sobre o mundo. Segundo Craik (1967), essa capacidade de modelagem é uma característica fundamental do pensamento e é responsável por proporcionar *insights* e a possibilidade de antecipar eventos, facilitando a adaptação comportamental frente a problemas. A teoria sugere que os seres humanos criam em suas mentes um "Modelo Mental" da realidade externa, que permite: experimentar diversas alternativas e escolher a melhor; preparar-se para situações futuras antes que ocorram; aplicar conhecimentos de eventos passados no presente e no futuro; responder de maneira eficaz e competente em situações de emergência. Para construir um modelo mental, Craik identifica três etapas essenciais após a observação da realidade externa: traduzir o mundo externo em palavras; deduzir uma assertiva a partir dessas observações; estabelecer uma conexão entre a assertiva e o mundo real. Esses modelos são, em essência, suposições sobre a realidade que influenciam o comportamento subsequente do indivíduo. A teoria destaca também que os modelos mentais facilitam a compreensão de conceitos abstratos ao torná-los visíveis e tangíveis mentalmente, reforçando assim sua aplicabilidade universal e sua utilidade prática (Da Costa, 2020).

Em sua concepção dos modelos mentais, Johnson-Laird (1983) adota conceitos menos abstratos de representação proposicional, aproximando-se mais da filosofia. Ele trata a teoria dos modelos mentais como uma representação mental de uma proposição expressa verbalmente. Em outras palavras, os modelos mentais são representações de objetos, eventos e estados das coisas, com a característica de serem indeterminadas, assim como as representações linguísticas (Johnson-Laird, 1983). De acordo com Johnson-Laird, na teoria dos modelos mentais, o pensamento depende de processos tácitos que são guiados por restrições, como o objetivo do pensador, quando aplicável, e conhecimentos e crenças relevantes.

Segundo Johnson-Laird (200), os modelos mentais têm a capacidade de representar relações espaciais, eventos, processos e as operações de sistemas

complexos. Eles podem ser utilizados para produzir inferências indutivas e dedutivas. Laird apresenta três suposições principais que distinguem os modelos mentais de outras formas de representação mental, como representações sintáticas de forma lógica e redes semânticas. Essas suposições são fundamentais na teoria dos modelos mentais:

- Cada modelo mental representa uma possibilidade, ou seja, ele captura o que é comum às diferentes maneiras pelas quais essa possibilidade pode ocorrer. Assim como um diagrama, um modelo é icônico, no sentido de que suas partes correspondem às partes do objeto que está sendo representado, e sua estrutura corresponde à estrutura da possibilidade em questão.

- O princípio da verdade: os modelos mentais representam o que é verdadeiro de acordo com as premissas, mas não representam automaticamente o que é falso. Ou seja, eles enfocam o que é consistente com as informações fornecidas.

- O raciocínio dedutivo depende dos modelos mentais: Se uma conclusão é válida em todos os modelos das premissas, ou seja, não existem contraexemplos, ela é necessária dadas as premissas. Se for válida em uma proporção de modelos, sua probabilidade é igual a essa proporção, desde que os modelos representem alternativas equiprováveis. Se for válida em pelo menos um modelo, é possível dadas as premissas. E se não se aplica a nenhum dos modelos, é impossível dadas as premissas.

A teoria dos modelos mentais de Johnson-Laird unifica o raciocínio dedutivo sobre necessidade, probabilidade e possibilidade. Segundo essa teoria, as pessoas não confiam automaticamente em regras formais de inferência, mas sim em seus modelos mentais, que são construídos com base na compreensão das premissas e no conhecimento geral. Dentre os três princípios mencionados, o princípio da verdade é considerado fundamental na teoria dos modelos mentais. Esse princípio afirma que "os raciocinadores representam apenas a informação mínima necessária em modelos explícitos, focando especificamente nas informações que são verdadeiras" (Johnson-

Laird & Savary, 1999).

Os modelos mentais desempenham um papel fundamental ao representar premissas e permitem o raciocínio sem depender exclusivamente da lógica formal. No entanto, a busca por interpretações alternativas requer uma representação independente das premissas, uma representação proposicional.

Johnson-Laird define os modelos mentais como representações de alto nível que estão no cerne psicológico da compreensão. Compreender algo significa ter um modelo mental ou uma representação de trabalho desse algo. Esses modelos são adquiridos por meio de observação, instrução ou inferência.

Norman (2014) destaca que uma característica dos modelos mentais é que eles são "não-científicos", ou seja, as pessoas tendem a manter padrões de comportamento "supersticiosos", mesmo quando sabem que não são necessários. Os modelos mentais de uma pessoa refletem suas crenças sobre o sistema físico representado, adquiridas por observação, instrução ou inferência.

Além disso, a capacidade das pessoas de executar tarefas e aprender depende dos conceitos associados que elas trazem consigo. Ao interagir com o ambiente, com outras pessoas e com artefatos tecnológicos, as pessoas formam modelos mentais internos de si mesmas e das coisas com as quais estão interagindo. Esses modelos mentais fornecem poder preditivo e explicativo para entender a interação.

3.2 Natureza dos Modelos Mentais e o Viés da Crença

Uma estrutura que pode ser utilizada para explicar o viés da crença no raciocínio humano é a Teoria dos Modelos Mentais, desenvolvida por Johnson-Laird (1983) e discutida por Oakhill e Johnson-Laird (1985), como aponta Torrens (1999). O viés da crença é um fenômeno específico em que as pessoas tendem a aceitar conclusões que estão alinhadas com suas crenças, independentemente da validade real dessas conclusões.

No processo de raciocínio dedutivo, os indivíduos constroem um modelo inicial, quase imaginário, das relações entre as premissas. No entanto, um

raciocinador prudente consideraria a existência de outros modelos que também fossem consistentes com as premissas. A partir disso, ele tiraria uma conclusão final que fosse consistente com todos os modelos gerados.

Como forma de estudar a teoria dos modelos mentais em relação ao viés da crença, Torrens (1999) conduziu um experimento com o objetivo de investigar as diferenças individuais nesse viés, que se refere à tendência de aceitar conclusões com base em sua plausibilidade em vez de sua validade lógica. De acordo com Torrens (1999), a maioria das pesquisas se concentra nos efeitos da crença no pensamento lógico, porém poucos estudos examinam se características individuais podem contribuir para esses efeitos.

Mesmo que a teoria dos modelos mentais seja a que parece ter uma melhor capacidade para representar o raciocínio humano, ainda existem algumas lacunas importantes que precisam ser preenchidas para que possamos considerá-los computáveis. Uma dessas lacunas é que não há um único modelo mental para um determinado estado de coisas, e não há uma determinação conceitual precisa do modelo. Na teoria, os modelos mentais são representações de alto nível mais abstratas, sem uma representação específica. Cada pessoa constrói seu próprio modelo mental do estado das coisas, podendo haver vários modelos, mesmo que apenas um deles represente de maneira ótima esse estado de coisas. Cada modelo mental é uma representação analógica desse estado de coisas e, inversamente, cada representação analógica corresponde a um modelo mental.

Embora a teoria dos modelos mentais seja uma abordagem amplamente utilizada para explorar o pensamento lógico, uma crítica a essa teoria é a falta de uma notação claramente definida. A ausência de uma notação específica impede que os modelos mentais sejam descritos de forma precisa e representados como procedimentos efetivos que possam ser executados por uma máquina. Além disso, não foram encontrados estudos que associem diretamente os modelos mentais à representação de conceitos morais.

Capítulo 4: Da Inteligência Artificial e da representação computacional do conhecimento

Este capítulo dedica-se à análise dos fundamentos e das técnicas avançadas no campo da IA, com ênfase especial na representação computacional do conhecimento. Iniciaremos com uma exploração dos conceitos básicos e do funcionamento geral da IA estabelecendo um entendimento inicial sobre como as máquinas processam e simulam a inteligência humana. Avançaremos para uma seção de discussão sobre a aprendizagem de máquina e a abordagem conexionista, que utiliza redes neurais para emular a capacidade de aprendizado do cérebro humano. Na seção seguinte serão abordados conceitos de *deep learning* ou aprendizagem profunda, uma técnica poderosa que tem revolucionado a capacidade das máquinas de reconhecer padrões complexos e tomar decisões autônomas. Além disso, abordaremos a modelagem cognitiva e simbólica, explorando como os sistemas de IA representam e manipulam simbolicamente o conhecimento humano.

Este capítulo também se dedica a uma discussão crítica sobre a representação do conhecimento moral humano em sistemas inteligentes, um desafio emergente na interseção de ética e tecnologia. Finalmente, investigaremos o raciocínio utilizando a teoria da lógica mental, que tenta compreender e replicar os processos de pensamento humanos dentro de sistemas computacionais. Ao longo deste capítulo, buscamos oferecer uma visão compreensiva das diversas facetas da IA e sua capacidade de modelar complexidades intelectuais e morais, destacando tanto as potencialidades quanto os limites da tecnologia atual.

4.1. Conceitos e funcionamento da Inteligência Artificial (IA)

Não há uma definição consensual do termo Inteligência Artificial. John McCarthy, em 1955 (Mccarthy et al., 2006), criou o termo "inteligência artificial" e o definiu como "a ciência e engenharia de fazer máquinas inteligentes". Russel, Norvig e Davis (2010) afirmam que a IA abrange vários subcampos, desde áreas de uso geral, como percepção e raciocínio lógico, até tarefas específicas como jogar xadrez, provar teoremas matemáticos, escrever poesia e diagnosticar doenças. Segundo esses autores, cientistas de outros campos estão gradualmente se voltando para a

inteligência artificial, pois é nessa área que encontram as ferramentas e o vocabulário para sistematizar e automatizar suas tarefas intelectuais. Da mesma forma, os pesquisadores em IA podem aplicar seus métodos a qualquer área de esforço intelectual humano. Nesse sentido, a Inteligência Artificial é entendida como um campo universal. Em uma revisão mais recente do conceito, Russell et al.(2022) conceituaram IA como o estudo de agentes que recebem percepções do ambiente e executam ações, onde cada agente implementa uma função que mapeia sequências de percepções em ações.

Russel, Norvig e Davis (2010) criaram uma classificação para a Inteligência Artificial, conforme apresentado no Quadro 1, que se divide em quatro dimensões: a) pensar humanamente; b) pensar racionalmente; c) agir humanamente; d) agir racionalmente. O Quadro 2 demonstra como essas quatro dimensões são reproduzidas em sistemas computacionais.

De acordo com Russel, Norvig e Davis (2010), essas definições variam ao longo de duas dimensões principais, em que os fatores da linha superior do quadro 1 estão preocupados com processos de pensamento e raciocínio, e os fatores da linha inferior com o comportamento. As definições à esquerda medem o sucesso em termos de desempenho humano, enquanto as definições à direita medem em relação a um conceito ideal de inteligência, a racionalidade. Como era de se esperar, existe uma tensão entre abordagens centradas em seres humanos e abordagens centradas na racionalidade, onde uma abordagem centrada no ser humano deve ser uma ciência empírica, envolvendo hipóteses e confirmação experimental, enquanto uma abordagem racionalista deve envolver uma combinação de matemática e engenharia. Portanto, com base nesse entendimento, a IA pode assumir diferentes formas, como a IA simbólica, o aprendizado de máquina e a IA conexionista (LIAO, 2020).

Quadro 1: *Algumas definições de IA. Elas estão organizadas em quatro categorias.*

Pensar Humanamente	Pensar Racionalmente
•"O novo e emocionante esforço para fazer os computadores pensarem... máquinas com mentes, no sentido pleno e literal" (Haugeland, 1985); •"A automação de] atividades que associamos ao pensamento humano, atividades como tomada de decisão, resolução de problemas, aprendizagem ..."(Bellman, 1978);	•“O estudo das faculdades mentais através do uso de modelos computacionais" (Charniak e McDermott, 1985); •"O estudo da computação que possibilitam perceber, raciocinar e agir" (Winston, 1992);
Agir Humanamente	Agir Racionalmente
•"A arte de criar máquinas que executam funções que requerem inteligência quando realizadas por pessoas" (Kurzweil, 1990); •"O estudo de como fazer computadores fazerem coisas nas quais, no momento, as pessoas são melhores" (Rich and Knight, 1991);	•"Um campo de estudo que busca explicar e emular comportamentos inteligentes em termos de processos computacionais" (Schalkoff, 1990); •"O ramo da ciência da computação que se preocupa com a automação do comportamento inteligente" (Luger e Stubblefield, 1993).

Fonte: Traduzido pela autora de Russel, Norvig e Davis (2010)

Cada uma dessas áreas da inteligência artificial possui diversas aplicações práticas em diferentes segmentos, mas todas ainda são bastante limitadas, especialmente quando nos deparamos com situações não previstas. Essas situações podem resultar em falhas sem consequências significativas ou em falhas graves, dependendo do sistema em que estão sendo aplicadas.

A IA simbólica é a coleção de métodos de pesquisa em inteligência artificial que se baseiam em representações simbólicas de alto nível de problemas, lógica e pesquisa. Ela está relacionada à forma como os seres humanos raciocinam e se popularizou com o surgimento dos Sistemas Especialistas. A IA simbólica tenta representar funções como pensar, aprender e resolver problemas por meio de raciocínio simbólico e lógica. Em particular, esses sistemas utilizam uma série de regras e declarações "se-então" explicitamente programadas para estabelecer as relações entre as entradas e saídas.

Por ser um sistema de raciocínio baseado em regras, a IA Simbólica também permite a visualização da lógica por trás do sistema, o que pode ser analisado e solucionado, se necessário. No entanto, a limitação da IA simbólica está na dificuldade de revisar as regras depois de codificadas para o sistema computacional (LIAO, 2020).

Quadro 2- Representação Computacional das categorias de Russel e Norvig (1995).

	Humano	Racional
Pensamento	Sistemas que pensam como humanos	Sistemas que pensam racionalmente
Comportamento	Sistemas que agem como humanos	Sistemas que agem racionalmente

Fonte: Interpretado pela autora de Russel e Norvig (1995).

4.2. Aprendizado de Máquina

O aprendizado de máquina utiliza algoritmos para aprender com dados, sem a necessidade de programação explícita. Dentro do conceito de aprendizado de máquina, podemos distinguir três tipos principais: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço (LIAO, 2020).

No aprendizado supervisionado, um algoritmo tem como objetivo aprender uma função que se aproxime melhor da relação entre entrada e saída nos dados. Para

isso, o algoritmo é treinado em um conjunto de dados de treinamento no qual as respostas corretas para determinados dados são conhecidas e os dados são rotulados de acordo. Dessa forma, o algoritmo pode utilizar as informações rotuladas para aprender a relação entre as entradas e as saídas. Uma vez que o algoritmo é treinado adequadamente, ele é capaz de aplicar o que aprendeu para prever a resposta correta em diferentes conjuntos de dados (LIAO, 2020).

No aprendizado não supervisionado, um determinado conjunto de dados não é rotulado e o algoritmo tem como objetivo encontrar padrões, estruturas ou agrupamentos nos dados por conta própria. Um exemplo de aprendizado não supervisionado é o aprendizado de regras de associação, no qual o algoritmo tenta descobrir regras que descrevam grandes porções dos dados. Para ilustrar esse tipo de aprendizado, podemos imaginar uma caixa contendo imagens de carros e motocicletas que não foram rotuladas ou classificadas (LIAO, 2020).

No aprendizado por reforço, o algoritmo busca aprender por meio da experiência. Nesse caso, o algoritmo é recompensado quando tem sucesso em uma tarefa e/ou punido quando falha. Através de tentativa e erro, o algoritmo se esforça para maximizar os resultados a longo prazo (LIAO, 2020).

4.3 Abordagem Conexionista na IA

A IA também pode assumir a forma conexionista, conhecida como aprendizagem profunda ou *deep learning*, que se baseia na simulação dos componentes do cérebro (modelagem da inteligência humana) e está enraizada no campo das redes neurais, também conhecido como o subcampo da inteligência artificial chamado "conexionismo" (GOODFELLOW et al., 2016).

A abordagem conexionista é baseada na ideia de que o comportamento inteligente pode ser entendido e modelado pela simulação da forma como os neurônios no cérebro humano processam informações. O bloco de construção fundamental desta abordagem é a rede neural artificial, que é um modelo computacional inspirado na estrutura e funcionamento de redes neurais biológicas (GOODFELLOW et al., 2016).

Os princípios do conexionismo e suas implicações para a inteligência artificial são discutidos por Russel, Norvig e Davis (2010), onde eles estabelecem que os sistemas conexionistas, em vez de dependerem apenas de programação explícita e manipulação simbólica, buscam aprender com os dados por meio do ajuste de pesos numéricos associados às conexões entre neurônios artificiais. Esse processo de aprendizagem é frequentemente referido como treinamento ou ajuste da rede neural.

De acordo com a abordagem conexionista, o aprendizado em redes neurais artificiais envolve o ajuste dos pesos da conexão com base em uma determinada entrada e na saída desejada. A rede processa iterativamente exemplos de treinamento e, por meio de um processo de minimização de erros, melhora gradualmente sua capacidade de produzir a saída correta para uma determinada entrada.

Os autores também discutem os benefícios dos sistemas conexionistas, como sua capacidade de aprender com a experiência, generalizar a partir de exemplos e exibir robustez e tolerância a falhas. Eles explicam que os modelos conexionistas podem capturar padrões e relacionamentos complexos nos dados, permitindo que eles resolvam problemas difíceis de resolver usando abordagens simbólicas tradicionais.

Além disso, Russel, Norvig e Davis (2010) destacam que os modelos conexionistas podem ser aplicados em diversas tarefas de IA, como reconhecimento de padrões, classificação, regressão, controle, reconhecimento de imagem, processamento de linguagem natural e robótica.

No geral, a abordagem conexionista descrita por Russell e Norvig enfatiza a importância das redes neurais e do aprendizado a partir de dados na construção de sistemas inteligentes.

4.4 *Deep-learning* ou Aprendizagem profunda

O aprendizado profundo, também conhecido como aprendizagem profunda, é um subcampo do aprendizado de máquina que está intimamente relacionado à abordagem conexionista descrita por Russel, Norvig e Davis (2010). O aprendizado

profundo se baseia nos princípios do conexionismo, utilizando redes neurais artificiais com várias camadas, daí o termo "profundo".

A aprendizagem profunda é um campo de estudo que busca capacitar máquinas a realizar tarefas inteligentes que antes eram realizadas exclusivamente por humanos. Atualmente, é um dos principais paradigmas de aprendizado, em que o software é capaz de inferir padrões automaticamente a partir dos dados disponíveis. No entanto, o aprendizado de máquina é tão bom quanto os dados com os quais ele é treinado. Se um algoritmo de aprendizado de máquina for treinado com dados inadequados ou imprecisos, ele fará previsões ruins, mesmo que tenha sido bem projetado.

Porém, Russel, Norvig e Davis (2010) apontam que se afirmarmos que um determinado programa pensa como um humano, devemos ter alguma maneira de determinar como os humanos pensam. Portanto, é necessário investigar o funcionamento real da mente humana, e existem duas maneiras de fazer isso: através da introspecção, tentando capturar nossos próprios pensamentos à medida que ocorrem, ou por meio de experimentos psicológicos.

Ou seja, uma vez que tenhamos uma teoria da mente suficientemente precisa, torna-se possível expressar essa teoria em um programa de computador. Consequentemente, um dos recursos para estabelecer uma teoria da mente pode ser encontrado na modelagem cognitiva simbólica, que é uma teoria da cognição humana expressa por meio de programas de computador funcionais.

Portanto, um modelo cognitivo busca ser uma explicação de como algum aspecto da cognição é realizado por um conjunto de processos computacionais primitivos. Segundo Marcus; Davis (2019), um desafio fundamental para a IA atualmente é encontrar um equilíbrio comparável entre mecanismos que capturam verdades abstratas e mecanismos que lidam com o mundo das exceções. Alcançar uma inteligência ampla exigirá que reunamos muitas ferramentas diferentes, algumas antigas e outras novas, de formas que ainda não descobrimos.

4.5 Modelagem cognitiva, simbólica

A modelagem desempenha um papel fundamental na construção científica e cognitiva, conforme destacado na enciclopédia "*The MIT Encyclopedia of the Cognitive Sciences*" (Wilson, 1999). Os modelos cognitivos simbólicos são teorias que visam descrever a cognição humana por meio de programas de computador funcionais. Esses modelos buscam explicar como um aspecto específico do processo cognitivo é realizado por meio de processos computacionais primitivos. Ao realizar uma tarefa cognitiva específica ou uma classe de tarefas, um modelo produz um comportamento que gera previsões passíveis de serem comparadas com dados de desempenho humano.

É fundamental fazer uma distinção entre as questões relacionadas à psicologia cognitiva e as questões computacionais. A modelagem cognitiva aborda questões científicas no âmbito da psicologia cognitiva, enquanto as técnicas computacionais são frequentemente derivadas da inteligência artificial. Através da modelagem cognitiva, é possível desenvolver uma teoria do comportamento humano em uma tarefa específica, o que viabiliza a criação de um artefato computacional capaz de executar essa tarefa.

Um modelo cognitivo, representado como um programa de computador, é uma descrição precisa dos processos envolvidos e de como eles se desenvolvem ao longo do tempo para realizar uma tarefa específica. Nesse contexto, a IA simbólica tem como objetivo aproximar o modelo cognitivo do modelo computacional, refletindo a maneira como os seres humanos processam informações e armazenam os símbolos necessários para compreender o mundo e as relações entre eles. Esses símbolos podem incluir palavras, cores, formas, sons e linguagem, sendo elementos essenciais na construção de modelos mentais.

Os modelos mentais são estruturas cognitivas concebidas para fundamentar o raciocínio, a tomada de decisão e, em certa medida, também influenciar os comportamentos. Esses modelos são construídos de forma individual, a partir das experiências pessoais de vida, percepções e compreensões do mundo, fornecendo mecanismos pelos quais novas informações são filtradas e armazenadas. Eles desempenham um papel fundamental na organização e interpretação das

informações, permitindo que sejam integradas ao conhecimento existente e utilizadas para orientar as ações e escolhas de uma pessoa.

A representação do conhecimento moral, assim como a representação do conhecimento em geral, é uma área interdisciplinar que demanda embasamento teórico e novas discussões. Envolve a utilização de diferentes métodos para representar o conhecimento moral, abrangendo uma variedade de elementos. Esses elementos incluem teorias éticas, que são estruturas sistemáticas para pensar sobre questões morais, como o consequencialismo, a deontologia e a ética da virtude. Além disso, envolvem princípios morais, que são regras ou diretrizes gerais utilizadas para determinar o que é certo ou errado, como o princípio de não causar danos ou o princípio da justiça. As regras morais também desempenham um papel importante, sendo prescrições específicas de como agir em determinadas situações, como "Não roube" ou "Não mate". Os julgamentos morais são avaliações de ações ou situações específicas como moralmente corretas ou incorretas, como o entendimento de que "É errado colar em uma prova". Além disso, as virtudes e vícios morais são características dos indivíduos, como honestidade ou egoísmo, geralmente consideradas moralmente boas ou más. As emoções morais, como empatia, culpa e vergonha, são frequentemente consideradas relevantes no contexto moral. As narrativas morais, como histórias ou exemplos, são utilizadas para ilustrar princípios ou julgamentos morais e apoiar o raciocínio moral. Por fim, as crenças morais representam conjuntos de crenças, costumes, valores e normas que orientam a ação e determinam o que é considerado correto e incorreto, tanto a nível individual quanto em grupos sociais. A complexidade desses elementos demanda uma abordagem cuidadosa e integrada para a representação do conhecimento moral.

4.6 Discussão sobre representação do conhecimento moral em sistemas computacionais

A representação do conhecimento humano em sistemas computacionais e o cálculo das consequências lógicas a partir desse conhecimento são desenvolvidos pela lógica. Através da dedução lógica, esses sistemas são capazes de inferir conhecimento implícito a partir das informações fornecidas. A representação formal

do conhecimento desempenha, portanto, um papel crucial na construção de sistemas computacionais inteligentes. Além disso, essas tecnologias também são amplamente utilizadas em aplicações que envolvem a criação de extensas bases de conhecimento, uma vez que o raciocínio lógico pode ajudar a evitar redundâncias e detectar erros. Nesta seção, nosso objetivo é apresentar as diferentes formas e características pelas quais o conhecimento pode ser representado no computador, bem como discutir como ele pode ser processado e interpretado de maneira formal

Ao compararmos algoritmos de aprendizagem profunda com estudos sobre a cognição humana, observamos que os seres humanos são capazes de aprender e fazer previsões com base em um número muito menor de exemplos, mesmo que esses exemplos sejam ruidosos e escassos, em comparação com os algoritmos. No entanto, os estudos também revelam que os seres humanos são capazes de justificar internamente suas decisões morais e articular razões para elas. Diante dessa diferença entre o raciocínio humano e os atuais algoritmos de aprendizado de máquina, surge a seguinte pergunta: como podemos utilizar as teorias mais recentes da ciência cognitiva para projetar inteligência artificial com a capacidade de aprender valores morais através de interações limitadas com humanos e tomar decisões com processos explicáveis? (KIM et al., 2018)

A modelagem do pensamento humano é complexa, uma vez que muitas das realizações humanas consistem em avaliações intuitivas com conceitos abstratos, que representam sentimentos subjetivos e não podem ser calculados ou enumerados. O conhecimento é complexo e multifacetado, portanto, apenas parcialmente compreensível. Lidar e calcular apenas com números não é suficiente para obter uma compreensão emocional das dimensões, do tempo e da extensão, o que é frequentemente essencial para avaliar corretamente os fatos situacionais concretos dentro de um contexto. Para que humanos e máquinas possam se comunicar, Dengel (2012) explica que a mensagem trocada deve representar a informação desejada, uma vez que a linguagem, escrita ou falada, é unidimensional, mas nosso ambiente é multidimensional e possui múltiplas camadas em seu contexto espacial e temporal de um evento ou situação. Portanto, devem ser utilizados formalismos que reflitam as circunstâncias e a estrutura de um evento ou situação, trazendo a situação para o contexto, o que permite a compreensão pela máquina.

A maioria das abordagens modernas para a representação do conhecimento baseia-se na lógica formal, em que a lógica descritiva e os formalismos baseados em regras são considerados as famílias linguísticas mais importantes. Nas aplicações práticas, as bases de conhecimento desempenham um papel fundamental na construção de ontologias, estabelecendo assim uma estreita relação entre a representação do conhecimento e outras áreas de pesquisa, como as tecnologias semânticas. Embora a pesquisa teórica em lógica formal seja a base da representação do conhecimento, esse campo também engloba várias questões aplicadas, como o desenvolvimento de ferramentas eficientes para o raciocínio automático. (BRACHMAN et al., 2004)

Uma das tentativas de representar o conhecimento humano por meio de um computador foi apresentada por John Searle em 1980, quando ele desenvolveu o experimento do pensamento conhecido como "quarto chinês". Nesse experimento, Searle argumentou que, embora um computador possa manipular os símbolos fornecidos, a máquina não atribui significados a esses símbolos. Ao realizar esse experimento, Searle pretendia demonstrar que os sistemas computacionais são limitados a operar apenas na sintaxe do programa implementado, sem compreender a semântica real por trás dos símbolos (SEARLE, 1980).

Para compreender o experimento do quarto chinês, é necessário considerar a visão de Wittgenstein no Tractatus, segundo a qual a verdade ou falsidade de uma proposição é determinada pela correspondência entre as palavras na proposição e os objetos no mundo. Portanto, deve haver uma identidade entre a estrutura das coisas, a estrutura do pensamento e a estrutura da linguagem. Dessa forma, podemos considerar que o processo dedutivo por si só não é totalmente capaz de representar o raciocínio humano.

Nesse contexto, também é relevante abordar o pensamento de Quine (2011), que questiona ou rejeita a noção de significado, pois isso pode levar a uma compreensão de um mundo em que existe apenas uma linguagem e nada a que a linguagem se refere. Quine argumenta que um objeto referido, nomeado por um termo singular ou denotado por um termo geral, pode ser qualquer coisa, ou seja, o significado de uma expressão é a ideia expressa. Essas considerações podem subsidiar a interpretação do experimento do quarto chinês, relacionando a noção

sintática à noção de objeto e a noção semântica à noção de significado.

Outra teoria recente no campo da ciência cognitiva, proposta por Kleiman-Weiner, Saxe e Tenenbaum (2017), parte do princípio de que os seres humanos aprendem a tomar decisões éticas ao adquirir princípios morais abstratos por meio da observação e interação com outros indivíduos em seu ambiente. Essa teoria descreve a decisão ética como uma escolha que busca maximizar a utilidade entre um conjunto de resultados, cujos valores são calculados com base em pesos atribuídos a conceitos morais abstratos, tais como "parentesco" ou "reciprocidade". Além disso, considerando a dinâmica dos indivíduos e sua participação em um grupo, essa estrutura explica como as preferências morais de um indivíduo e suas ações resultantes contribuem para o desenvolvimento de princípios morais compartilhados pelo grupo (ou seja, normas do grupo). A principal contribuição dessa teoria é formalizar um conjunto mínimo de capacidades cognitivas que as pessoas podem utilizar para resolver problemas de aprendizagem, baseadas em três componentes:

a) Um cálculo de utilidade abstrata e recursiva, onde as teorias morais podem ser formalizadas como valores ou pesos que um agente atribui a um conjunto de princípios abstratos. Esses princípios são utilizados para fatorar as funções de utilidade de outros agentes em suas próprias tomadas de decisão e julgamentos baseados em utilidade.

b) Inferência bayesiana hierárquica, que permite inferir de forma rápida e confiável os pesos que outros agentes atribuem a esses princípios morais abstratos através da observação de seus comportamentos. Através desse mecanismo, é possível realizar a aprendizagem moral no nível dos valores, em vez de apenas imitar comportamentos.

c) Aprendizagem por alinhamento de valor, que envolve a definição de valores próprios guiados por meta-valores ou princípios. Esses meta-valores buscam alinhar as teorias morais de uma pessoa com as dos outros ("valorizamos os valores daqueles que valorizamos"), além de serem coerentes com os próprios apegos e sentimentos. Esse componente aborda a dinâmica da aprendizagem moral, explorando quando e como somos motivados a adquirir ou mudar nossos próprios

valores, mesmo após inferir valores morais dos outros.

Essa teoria caracteriza a decisão ética como uma escolha que busca maximizar a utilidade em relação a um conjunto de resultados, cujos valores são calculados com base nos pesos atribuídos pelos indivíduos a conceitos morais abstratos, como "parentesco" ou "relação recíproca". Além disso, considerando a dinâmica dos indivíduos e sua participação em um grupo, essa estrutura explica como as preferências morais de um indivíduo e as ações resultantes delas contribuem para o desenvolvimento de princípios morais compartilhados pelo grupo, ou seja, as normas do grupo.

Para testar a teoria de Kleiman-Weiner, Saxe e Tenenbaum (2017), Kim et al. (2018) expandiram a estrutura apresentada para explorar um modelo computacional da mente humana em dilemas morais com decisões binárias. Eles caracterizaram a tomada de decisão em dilemas morais como uma função de utilidade que calcula os *trade-offs* de valores percebidos pelos humanos nas escolhas do dilema. Esses valores são representados pelos pesos atribuídos pelos humanos às dimensões abstratas do dilema, chamados pelos autores de "pesos de princípios morais". Além disso, representaram um agente individual como membro de um grupo que compartilha princípios morais semelhantes. Esse compartilhamento dos princípios morais do grupo como um agregado dá origem à norma do grupo, demonstrando como a inferência bayesiana hierárquica pode fornecer um mecanismo poderoso para inferir rapidamente princípios morais individuais e a norma do grupo, mesmo com dados limitados e ruidosos.

Em um estudo recente, Bandy (2021) realizou uma revisão sistemática que analisou 62 estudos que auditaram algoritmos direcionados ao público. Nessa revisão, Bandy identificou quatro tipos de comportamentos problemáticos em algoritmos que utilizam inteligência artificial: discriminação, distorção, exploração e erro de julgamento. A maioria dos estudos revisados concentrou-se na análise da discriminação e distorção.

Existem vários tipos de erros que podem ocorrer em algoritmos usados para tomar decisões morais, segundo Bandy (2021), incluindo:

a) Viés: ocorre quando o algoritmo é treinado em dados tendenciosos, o

que pode levar a decisões tendenciosas. Por exemplo, se um algoritmo usado para prever a reincidência (a probabilidade de um criminoso condenado cometer outro crime) for treinado em dados de um sistema prisional que aprisiona desproporcionalmente indivíduos negros, pode ser mais provável prever a reincidência de indivíduos negros, mesmo que eles não sejam mais propensos à reincidência do que indivíduos de outras raças.

b) *Overfitting*: ocorre quando o algoritmo se ajusta em excesso aos dados de treinamento. Isso significa que o modelo se adapta muito bem aos dados de treinamento, mas seu desempenho é ruim quando aplicado a dados de teste. O sistema conhece tão bem os dados de treinamento que acaba aplicando regras específicas apenas a esses dados.

c) Excesso de confiança: ocorre quando o algoritmo é excessivamente confiante em suas previsões, levando-o a tomar decisões arriscadas. Isso é especialmente problemático em situações em que as consequências de uma decisão errada são graves.

d) Falta de transparência: ocorre quando o algoritmo é difícil de entender, tornando difícil identificar erros ou avaliar seu desempenho. Isso é um problema quando o algoritmo é usado para tomar decisões importantes, pois pode ser difícil determinar se ele está agindo no melhor interesse de todas as partes envolvidas.

e) Falta de responsabilidade: ocorre quando o algoritmo é visto como uma "caixa preta" e é difícil atribuir responsabilidade por quaisquer erros cometidos. Isso pode dificultar a responsabilização de qualquer pessoa por danos causados pelas decisões do algoritmo.

No geral, o principal desafio para a tomada de decisões algorítmicas em situações moralmente ou eticamente desafiadoras é garantir que o algoritmo seja transparente, responsável, explicável e imparcial, a fim de evitar danos no processo. Portanto, é necessário estabelecer um método de representação do conhecimento moral que leve em consideração uma tomada de decisão ótima e, em seguida, especificar um algoritmo computacional.

No campo da representação computacional do raciocínio humano, é importante considerar as limitações que surgem quando lidamos com máquinas. Segundo Johnson-Laird (2010), embora as máquinas possuam um potencial quase ilimitado de aprendizado, o raciocínio baseado em conectivos sentenciais, como "se" e "ou", torna-se computacionalmente intratável. Conforme aumentamos o número de proposições elementares distintas nas inferências, o processamento exigido pelo raciocínio excede a capacidade de qualquer dispositivo computacional finito, incluindo o cérebro humano. Isso nos leva a questionar se é possível representar o raciocínio e o conhecimento moral de maneira computável.

Nesse contexto, surge a incerteza sobre a viabilidade de representar a complexidade do raciocínio humano e a natureza subjetiva do conhecimento moral através de abordagens computacionais. Embora as máquinas tenham demonstrado um grande potencial de aprendizado, a questão de como incorporar os aspectos qualitativos e sutis do raciocínio humano e do julgamento moral permanece em aberto. Avançar nessa área requer um aprofundamento das teorias e uma busca por métodos que possibilitem uma representação mais abrangente e precisa do conhecimento humano, levando em consideração os desafios computacionais intrínsecos ao raciocínio lógico e moral.

Ao explorarmos essa intersecção entre a capacidade de processamento das máquinas e a complexidade do pensamento humano, abrimos caminho para avanços significativos na inteligência artificial e na compreensão da natureza humana. A busca por métodos e abordagens que possam lidar com os desafios inerentes à representação do raciocínio humano e do conhecimento moral é fundamental para o desenvolvimento de sistemas mais inteligentes e éticos.

O estudo do raciocínio dedutivo humano através de hipóteses inferenciais tem sido objeto de investigação científica interdisciplinar. Durante muito tempo, a lógica inferencial foi a base para a programação de sistemas inteligentes. De acordo com Johnson-Laird (2010), havia um consenso na psicologia de que nossa capacidade de raciocínio dependia de uma lógica mental tácita, composta por regras formais de inferência semelhantes às de um cálculo lógico. No entanto, essa hipótese encontrou dificuldades que levaram a uma perspectiva alternativa: o raciocínio depende da consideração de possibilidades consistentes com o ponto de partida, que pode ser

uma percepção do mundo, um conjunto de afirmações, uma memória ou uma combinação desses elementos. Quando raciocinamos, podemos ter como objetivo obter conclusões válidas, mas existem muitas conclusões válidas que nunca alcançamos. Portanto, afirmar que o raciocínio se resume apenas à lógica é ignorar esse fato, como observado por Laird.

Essa visão alternativa destaca a importância de explorar a capacidade humana de raciocinar além das regras formais de inferência. Envolve considerar o processo criativo de vislumbrar possibilidades e explorar diferentes caminhos para o pensamento. Compreender como a mente humana realiza essas inferências além da lógica formal é crucial para a construção de sistemas mais inteligentes e para uma melhor compreensão dos mecanismos cognitivos envolvidos no raciocínio humano.

Os formalistas defendem a ideia de que as pessoas nascem com a capacidade de raciocínio lógico e possuem regras de inferência esquemáticas inatas. No entanto, os estudos de Gödel revelaram uma das fraquezas do formalismo ao abordar a questão da consistência em sistemas axiomáticos (Netto, 2011).

Além disso, os formalistas também consideram a possibilidade de informatizar toda a matemática, o que nos leva a questionar filosoficamente a capacidade dos computadores de pensar. Ao compreendermos o comportamento humano como orientado por um conjunto de julgamentos ou juízos, que determinam a interpretação da realidade e o valor das ações, e considerando que as teorias do conhecimento e da justificação dependem crucialmente da noção de verdade, surge a indagação sobre se a lógica é a ciência que nos aproxima melhor da representação da verdade.

A capacidade de raciocinar e utilizar a lógica é uma das características distintivas da inteligência humana. No entanto, é importante também considerar o viés de crença no processo de raciocínio. O viés de crença se refere à tendência de aceitar conclusões que acreditamos serem verdadeiras, independentemente de sua validade a partir das premissas. Em outras palavras, a avaliação da consistência lógica do argumento é influenciada pela credibilidade da conclusão (Wason & Johnson-Laird, 1972).

No que diz respeito à representação computacional do raciocínio humano, Johnson-Laird afirma que, em teoria, qualquer teoria científica pode ser modelada em

um programa de computador. No entanto, a questão crítica reside em saber se esse programa incorpora o fenômeno do pensamento humano. Johnson-Laird sugere que talvez seja possível alcançar isso por meio do desenvolvimento de uma teoria robusta do pensamento humano. No entanto, os programas existentes ainda não conseguem igualar o pensamento humano. Alguns desses programas são produtos da inteligência artificial, enquanto outros são projetados para modelar teorias do pensamento humano. Embora eles simulem processos de dedução, indução e criação, ainda não os incorporam completamente (Johnson-Laird, 2015).

4.7 Raciocínio utilizando a teoria da lógica mental

A teoria da lógica mental postula que as pessoas raciocinam construindo modelos do estado de coisas, formulando conclusões com base nesses modelos e procurando modelos alternativos que os refutem. A lógica mental desempenha um papel fundamental na integração de informações e na realização de inferências, uma vez que as pessoas estão frequentemente expostas a informações provenientes de várias fontes e precisam de uma lógica mental para integrar essas informações, conectá-las a informações previamente adquiridas e fazer inferências que vão além dessas informações. Enquanto as regras formais funcionam de forma sintática, os modelos mentais operam de forma semântica (Braine & O'Brien, 1998).

A hipótese de que o raciocínio depende de uma lógica mental postula duas etapas principais para realizar uma inferência dedutiva: a recuperação da forma lógica das premissas e o uso de regras formais para provar uma conclusão. Johnson-Laird (2010) apresenta o seguinte exemplo para ilustrar a compreensão da lógica mental:

Considere as seguintes premissas:

1. Ou o mercado tem um desempenho melhor, ou não poderei me aposentar.

2. Eu poderei me aposentar.

A primeira premissa é uma disjunção exclusiva, onde apenas uma das cláusulas pode ser verdadeira, mas não ambas. Como as teorias psicológicas não possuem regras para lidar com disjunções exclusivas, a primeira premissa é atribuída uma forma lógica que abrange uma disjunção inclusiva, permitindo que ambas as

cláusulas possam ser verdadeiras, com uma negação dessa situação:

1. A ou ~B & ~ (A & ~B).

2. B

De acordo com as regras formais, Johnson-Laird (2010) afirma que elas fornecem uma prova para A, correspondendo à conclusão:

3. O mercado tem um desempenho melhor.

O número de etapas em uma prova deve prever a dificuldade de uma inferência, e algumas evidências corroboram essa previsão. No entanto, Johnson-Laird (2010) destaca que existem desafios para a lógica mental, sendo um deles a etapa inicial de recuperação da forma lógica das afirmações. Para ilustrar esse problema, o autor utiliza o famoso conselho do Sr. Micawber, presente no romance "David Copperfield" de Dickens.

Considere as seguintes premissas:

1. Uma renda anual de vinte libras, despesas anuais dezenove libras, dezenove (xelins) e seis (pence), resultam em felicidade

2. Uma renda anual de vinte libras, despesas anuais de vinte libras e seis (pence), resultam em miséria.

Johnson-Laird (2010) aponta que não há algoritmo que possa recuperar a forma lógica de todas as afirmações do dia a dia. A dificuldade reside no fato de que a forma lógica necessária para a lógica mental não é apenas uma questão de sintaxe das sentenças, mas depende do conhecimento, como o fato de que dezenove libras e dezenove xelins e seis pence são menos do que vinte libras, e que felicidade e miséria são propriedades inconsistentes. Além disso, a forma lógica pode depender do conhecimento do contexto em que uma frase é proferida, como quando o falante aponta para coisas no mundo. Portanto, alguns lógicos duvidam que a forma lógica seja relevante para o raciocínio cotidiano, e sua extração pode depender, por sua vez, do próprio raciocínio, o que pode levar a uma regressão infinita (Johnson-Laird, 2010). Outras dificuldades também são encontradas na lógica mental, conforme apontado

por Johnson-Laird (2010):

a) Dadas as premissas do exemplo sobre aposentadoria, qual conclusão devemos tirar? A lógica produz apenas uma restrição. A conclusão deve ser válida, ou seja, se as premissas são verdadeiras, a conclusão também deve ser verdadeira. Portanto, a conclusão deve ser válida em todas as possibilidades em que as premissas são válidas. No entanto, a lógica produz um número infinito de inferências válidas a partir de qualquer conjunto de premissas. Muitas dessas conclusões válidas são tolas e a tolice dificilmente é racional. Assim, a lógica sozinha não pode caracterizar completamente o raciocínio racional. As teorias da lógica mental buscam evitar inferências tolas, mas ao mesmo tempo têm dificuldade em explicar como reconhecemos que uma inferência tola é válida.

b) Outra dificuldade para a lógica mental é que fazemos deduções válidas quando fatos brutos entram em conflito com elas. A lógica pode estabelecer tais inconsistências - na verdade, um método lógico as explora para produzir inferências válidas: negamos a conclusão a ser provada, a adicionamos às premissas e mostramos que o conjunto resultante de sentenças é inconsistente. No entanto, na lógica ortodoxa, qualquer conclusão decorre de uma contradição e, portanto, nunca é necessário retirar uma conclusão. A lógica é, portanto, monotônica: quanto mais premissas são adicionadas, maior é o número de conclusões válidas que podem ser tiradas de forma monotonicamente crescente.

c) Um outro problema para a lógica mental é que as manipulações de conteúdo afetam as escolhas dos indivíduos sobre quais casos refutam uma hipótese geral, em um fenômeno conhecido como "tarefa de seleção" (WASON et al., 1972). O desempenho nessa tarefa é passível de várias interpretações, incluindo a visão de que os indivíduos não se concentram em raciocínio dedutivo, mas sim em otimizar a quantidade de informações que provavelmente obterão a partir das evidências. No entanto, esses problemas levaram a uma concepção alternativa do

raciocínio humano.

Segundo Johnson-Laird (2010), os humanos não raciocinam de forma linear, pois tendem a inferir conclusões que entram em conflito com fatos brutos. Essa propensão, no entanto, é considerada racional, tanto que os teóricos desenvolveram sistemas de raciocínio não monotônicos. Ao raciocinarmos, buscamos conclusões que sejam verdadeiras ou, pelo menos, prováveis, com base nas premissas, e buscamos conclusões que sejam novas e que agreguem informações.

Portanto, ao raciocinarmos, não chegamos a uma conclusão que apenas repita uma premissa, que seja uma simples conjunção das premissas, ou que adicione uma alternativa às possibilidades mencionadas nas premissas, mesmo que esses tipos de conclusão sejam válidos. Em vez disso, buscamos uma relação ou propriedade que não tenha sido explicitamente afirmada nas premissas. Dependendo se essa relação se aplica a todos, à maioria ou a alguns dos modelos considerados, chegamos a uma conclusão sobre sua necessidade, probabilidade ou possibilidade. Para alcançar esse tipo de raciocínio, é necessário recorrer ao raciocínio baseado em modelos.

Capítulo 5: Representação lógica do conhecimento abstrato

Este capítulo aborda a representação lógica do conhecimento abstrato, explorando as interfaces complexas e multifacetadas entre lógica, conhecimento e crença. A investigação inicia-se com uma análise da lógica como fundamento para a representação estruturada do conhecimento. Prosseguimos com uma discussão sobre o raciocínio não monotônico, um tipo de raciocínio lógico que permite conclusões a serem retiradas mesmo na presença de conhecimento incompleto ou em evolução, refletindo assim a maneira flexível e adaptativa com que humanos pensam e agem. Adentraremos ainda mais na complexidade da lógica com a exploração da lógica não-monotônica, que desafia as premissas tradicionais da lógica clássica para lidar com incertezas e informações que mudam ao longo do tempo. Em sequência, a lógica modal será introduzida como uma ferramenta para modelar modalidades de possibilidade e necessidade.

Este capítulo pretende oferecer uma visão de como a lógica contribui para a modelagem e o entendimento de conhecimentos abstratos, fundamentais tanto para o desenvolvimento teórico quanto para aplicações práticas em áreas como inteligência artificial envolvendo ética e conhecimento abstrato. Através dessa análise, buscamos elucidar como sistemas formais de lógica podem ser aplicados para representar complexidades da moralidade humana para a decisão moral.

5.1 Lógica, conhecimento e crença

A capacidade de inferir conhecimento é um componente central da habilidade humana de resolver problemas. Com base em fatos existentes, somos capazes de concluir que certos eventos plausíveis ocorreram ou fazemos suposições ao interpretar as conexões entre os eventos. Na área de inteligência artificial, uma das principais preocupações é investigar e descrever formalmente essas técnicas de inferência. Para isso, a lógica é frequentemente utilizada como um formalismo adequado para representar o conhecimento em certas circunstâncias. A lógica matemática possui um poder significativo que permite abstrair o pensamento humano e construir bases de conhecimento a partir das quais é possível tirar conclusões ou derivar novos fatos. Por meio de sistemas lógicos, é possível modelar e representar

relações entre fatos, fazer deduções lógicas e estabelecer conexões precisas entre diferentes informações (RUSSELL, S., & NORVIG, P., 2016).

Entretanto, é importante reconhecer que a lógica possui suas limitações, especialmente quando lidamos com incerteza, ambiguidade e contextos complexos. Diferentes paradigmas e abordagens, como o aprendizado de máquina e a representação de conhecimento baseada em redes neurais, têm sido explorados para complementar a lógica na resolução de problemas mais desafiadores. Também é importante ressaltar que não existe uma lógica universal capaz de expressar todas as características de todos os problemas do mundo. Na prática, são criados diferentes sistemas lógicos, cada um com suas características particulares, para abordar diferentes tipos de problemas.

Quando se trata da representação do conhecimento moral, o processo envolve a representação do conhecimento abstrato, e tanto a lógica formal quanto as ontologias podem ser usadas para representar esse tipo de conhecimento. No entanto, esses recursos lógicos apenas podem expressar se algo é verdadeiro ou falso, o que pode ser problemático no caso do conhecimento abstrato, pois o raciocínio frequentemente envolve fatos que são verdadeiros na maioria dos casos, mas nem sempre. Um exemplo clássico que pode ilustrar esse contexto é:

a) “pássaros normalmente voam”;

Esta regra pode ser expressa na lógica padrão por:

b) “todos os pássaros voam”;

O que é inconsistente com o fato de que os pinguins não voam, ou

c) “todas as aves que não são pinguins e não avestruzes e voam”;

O que requer exceções à regra a ser especificada.

Portanto, é um exemplo amplamente utilizado para demonstrar as limitações da lógica tradicional na representação de conhecimento que envolve exceções, como no caso dos pinguins que são aves, mas não voam. A lógica padrão visa formalizar regras de inferência como essa sem mencionar explicitamente todas as suas

exceções. Há então a necessidade de analisar estudos que contemplam o uso da lógica para a representação do conhecimento abstrato, e identificar fundamentos importantes que envolvem essa representação. Alguns estudos como os de Fagin e Halpern, (1987), Stalnaker (2006), Baltag et al., (2017) e Wáng, (2021), discutem como o projeto semântico formal pode ser usado para esclarecer a relação entre conceitos epistêmicos e outros modais.

No estudo de Fagin e Halpern (1987), foram discutidas e estudadas novas lógicas para crença e conhecimento, considerando a limitação dos agentes em não serem logicamente oniscientes. O estudo propõe três lógicas distintas. A primeira lógica é uma extensão da lógica de Levesque, abordando a crença implícita e explícita, permitindo múltiplos agentes e crenças de nível superior, ou seja, crenças sobre crenças. A segunda lógica aborda explicitamente a "consciência", enfatizando que é necessário ter conhecimento prévio de um conceito antes de formar crenças sobre ele. Por fim, a terceira lógica propõe um modelo de "raciocínio local", considerando um agente como uma "sociedade de mentes", cada uma com suas próprias crenças, que podem se contradizer.

No entanto, os autores reconhecem que, nessas lógicas, o conjunto de crenças de um agente pode não conter todas as fórmulas válidas. Mesmo assim, eles consideram essas lógicas mais adequadas do que as tradicionais para modelar as crenças de seres humanos ou máquinas com capacidades de raciocínio limitadas. Uma limitação do estudo de Fagin e Halpern (1987) está na especificação de uma noção particular de consciência, uma vez que as condições da função de consciência devem ser determinadas por cada aplicação específica. Portanto, mais pesquisas são necessárias para encontrar funções de consciência úteis e naturais. Os autores também sugerem a combinação de consciência e tempo como uma direção promissora para modelar propriedades de aquisição de conhecimento, considerando versões quantificadas das lógicas propostas.

Segundo Stalnaker (2006), a estrutura semântica formal pode fornecer os recursos para construir modelos que ajudam a esclarecer a relação abstrata entre o conceito de conhecimento e outros conceitos, como crença, revisão de crença, causalidade e contrafactuais. Stalnaker (2006) utiliza uma semântica relacional baseada em modelos de Kripke que são reflexivos, transitivos e direcionados. Em seu

trabalho, ele analisa a relação entre conhecimento e crença, e desenvolve um sistema modal combinado para essas noções, com base nos axiomas derivados dessa análise. No entanto, o modelo de Stalnaker (2006) é adequado apenas quando as informações provêm de fontes discretas, mesmo que o agente não identifique ou distinga essas fontes. No entanto, o modelo não aborda os problemas abstratos que surgem quando a estrutura capturada pelo modelo se torna mais complexa.

De acordo com Baltag (2017), compreender a relação entre conhecimento e crença é uma questão central na epistemologia formal. A distinção entre crença e conhecimento e como um pode ser definido em termos do outro se tornou cada vez mais relevante. O autor destaca que esse problema tem sido abordado a partir de duas perspectivas opostas na literatura. Algumas propostas seguem a linha da crença verdadeira justificada como base do conhecimento, aceitando a prioridade conceitual da crença sobre o conhecimento. Nessa abordagem, parte-se de uma noção fraca de crença, que é pelo menos justificada e verdadeira, e busca-se fortalecer essa noção para definir o conhecimento, de modo que a noção definida de conhecimento não esteja sujeita a contraexemplos. Por outro lado, a segunda abordagem apresentada por Baltag (2017) desafia a "prioridade conceitual da crença sobre o conhecimento" e reverte a relação, dando prioridade ao conhecimento. Quando o conhecimento tem prioridade, outras atitudes, como a crença, devem ser explicáveis ou definíveis em termos dele, conforme apresentado por Stalnaker (2006)..

No trabalho de Wáng (2021), é apresentada uma lógica do conhecimento em um quadro no qual o conhecimento é tratado como uma forma de crença. O autor utiliza o axioma KD45 padrão de crença e define o conhecimento com base no conceito clássico de conhecimento como crença verdadeira justificada, que se relaciona naturalmente com os estudos sobre lógicas de evidência e justificação.

O axioma KD45 na lógica modal combina várias propriedades importantes dos operadores modais de necessidade (□) e possibilidade (◊). Ele é tipicamente usado em contextos em que a modalidade em questão é interpretada como "conhecimento" ou "crença". Aqui estão os componentes do axioma KD45 e o que eles representam:

K: Esse é o axioma da distributividade, que afirma que se é conhecido que uma implicação é verdadeira, então se a premissa é conhecida, a conclusão também deve

ser conhecida. Em termos formais, ele é expresso como:

Este axioma é básico para sistemas de lógica modal que envolvem raciocínio sobre conhecimento e crença.

D: Conhecido como o axioma da serialidade, ele afirma que se algo é necessariamente verdadeiro, então também é possível que seja verdadeiro. Isso elimina a possibilidade de um mundo onde uma proposição é necessariamente verdadeira, mas nunca possível.

Este axioma é usado em sistemas onde cada estado possível tem um sucessor (ou seja, não há "fim do mundo").

4: Este axioma afirma que se algo é conhecido, então é conhecido que é conhecido. Isso introduz uma forma de introspecção positiva, onde o agente é ciente de seus próprios conhecimentos:

5: Este axioma afirma que se algo não é conhecido, então é conhecido que não é conhecido. Isso introduz uma forma de introspecção negativa, permitindo ao agente reconhecer a ausência de certos conhecimentos:

Juntos, os axiomas K, D, 4 e 5 formam o sistema KD45, que é comumente aplicado no estudo de sistemas de crenças. Este sistema é útil para modelar agentes que são considerados introspectivos (conscientes de seus próprios conhecimentos e ignorâncias) e racionais em suas crenças, mas que ainda assim podem ter crenças falsas. Ele é especialmente relevante para teorias de crenças justificadas ou para ambientes onde as suposições devem ser consistentes e fechadas sob conhecimento.

De acordo com Wáng (2021), essa interpretação do conhecimento evita propriedades indesejadas da onisciência lógica, independentemente da escolha da lógica subjacente à crença. O autor demonstra a solidez e completude da axiomatização da lógica, bem como estuda os resultados de complexidade computacional relacionados aos problemas de verificação e adequação do modelo.

A interpretação do conhecimento refere-se à maneira como os indivíduos compreendem e atribuem significado às informações que recebem. Ela auxilia na evitação das propriedades indesejadas da onisciência lógica, evidenciando que os

seres humanos possuem habilidades cognitivas limitadas e que nosso conhecimento é sempre provisório e sujeito a revisões. Essa interpretação envolve um processo subjetivo de análise, avaliação e integração de informações na base de conhecimento existente.

Nos estudos mencionados, fica claro que todos eles optam pela interpretação do conhecimento como forma de evitar a onisciência lógica. O conceito de onisciência lógica pressupõe que um agente, em uma determinada situação, conhece todas as consequências lógicas das informações que possui ou é capaz de deduzir todas as possíveis consequências a partir de um conjunto de premissas. No entanto, essa é uma noção idealizada que raramente, ou nunca, pode ser alcançada na prática e não seria viável em cenários do mundo real.

A interpretação do conhecimento permite que os indivíduos reconheçam e aceitem a incerteza e a ambiguidade em sua compreensão do mundo. Ela também possibilita a existência de múltiplas interpretações válidas de um determinado conjunto de fatos, reconhecendo que essas interpretações podem mudar ao longo do tempo à medida que novas informações são adquiridas. Essa abordagem fornece uma visão mais realista da aquisição e uso do conhecimento, evitando as suposições problemáticas e limitações da onisciência lógica. As linguagens de lógica possuem características interessantes que as tornam uma boa escolha para representar o conhecimento. Jurafsky e Martin (2000) afirmam que as lógicas modais têm sido amplamente utilizadas na representação do conhecimento do senso comum, além de serem aplicadas na modelagem de crenças, representação do tempo e construção de cenários hipotéticos.

Dentro deste contexto, Juhos (2009) apresenta em seu trabalho o questionamento sobre o raciocínio com afirmações condicionais na forma "se p, então q". De acordo com o autor, apesar do raciocínio humano a partir de afirmações condicionais parecer intuitivo e fazer parte do nosso cotidiano, além de constituir uma importante ferramenta cognitiva, representa um quebra-cabeça conceitual. Em seu trabalho, o autor investiga como a solução da problemática das condicionais contribui para uma melhor compreensão da racionalidade humana, por meio da aplicação da teoria dos modelos utilizando a lógica modal como forma de representação.

Segundo o autor, embora as origens do estudo do raciocínio condicional estejam enraizadas na lógica proposicional, essa abordagem se mostra insuficiente para dar conta da pluralidade dos significados que o "se" assume na linguagem natural. Para abordar essa questão, Juhos dedicou-se à investigação da teoria dos modelos mentais, que abrange uma grande variedade de domínios. Entre os domínios da teoria dos modelos citados por Juhos (2009), destacam-se:

a) Raciocínio causal (Goldvarg & Johnson-Laird, 2001; Johnson-Laird & Goldvarg-Steingold, 2007)

b) Raciocínio deôntico (Bucciarelli & Johnson-Laird, 2005; Manktelow & Over, 1992; Quelhas & Byrne, 2003)

c) Raciocínio contrafactual (Byrne, 2005; Byrne & McEleney, 2000; Byrne & Tasso, 1999; Quelhas & Byrne, 2000)

d) Raciocínio modal (Bell & Johnson-Laird, 1998; Evans, Handley, Harper, & Johnson-Laird, 1999; Goldvarg & Johnson-Laird, 2000)

e) Raciocínio probabilístico (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999; Johnson-Laird & Savary, 1996)

f) Raciocínio temporal (Schaeken, Johnson-Laird, & d'Ydewalle, 1996; Vandierendonck, De Vooght, & Dierckx, 2000)

g) Raciocínio baseado em suposições (Byrne & Handley, 1997)

h) Detecção de inconsistências e sua resolução (Girotto, Johnson-Laird, Legrenzi, & Sonino, 2000)

i) Argumentação informal (Green, 2007)

j) Argumentação pragmática (Manktelow, Fairley, Kilpatrick, & Over, 2000; Sperber, Cara, & Girotto, 1995), entre outros.

k) Dentro deste contexto, passamos a discutir o uso da lógica modal como forma de representação do conhecimento moral e da motivação.

5.2 Raciocínio Não Monotônico

A escolha da lógica modal para trabalhar com a representação do conhecimento moral e da motivação moral pode ser justificada ao avaliar a questão do raciocínio de senso comum, que transita pelo mesmo universo e pode ser considerado não monotônico. Segundo Moore (1985), o raciocínio de senso comum é "não monotônico" porque muitas vezes o ser humano chega a conclusões com base em informações parciais, que podem ser revistas quando informações mais completas são recebidas posteriormente.

Portanto, o raciocínio não monotônico se tornou a principal ferramenta para o tratamento computacional do conhecimento moral abstrato. O autor ilustra a não monotonicidade por meio do seguinte exemplo:

> Se sabemos que o Piu-Piu é um pássaro, normalmente assumiremos, na ausência de evidências em contrário, que o Piu-Piu pode voar. Se, no entanto, descobrirmos mais tarde que Piu-Piu é um pinguim, retiraremos nossa suposição anterior. Se tentarmos modelar isso em um sistema formal, parece que temos uma situação na qual um teorema P é derivável de um conjunto de axiomas s, mas não é derivável de algum conjunto S' que é um superconjunto de s. O conjunto de teoremas, portanto, não aumentam monotonicamente com o conjunto de axiomas; portanto, esse tipo de raciocínio é considerado "não-monotônico" (Moore, 1985, P.80).

As lógicas padrão são sempre monotônicas, de acordo com Minsk (1974), porque suas regras de inferência tornam todo axioma permissivo. Ou seja, as regras de inferência são sempre da forma:

"P é um teorema se Q1..., Qn não são teoremas"

A inferência de que os pássaros podem voar é tratada por meio de uma regra que diz que:

Para qualquer X,

"X pode voar" é um teorema

se "X é um pássaro" é um teorema

"X não pode voar",

“voar" não é um teorema.

Se tudo o que nos dizem sobre Piu-Piu é que ele é um pássaro, não seremos capazes de derivar "Piu-Piu não pode voar"; consequentemente, " Piu-Piu pode voar" será inferido. Se nos disserem que “Piu-Piu é um pinguim” e já sabemos que nenhum pinguim pode voar, seremos capazes de derivar o fato de que “Piu-Piu não pode voar” e, portanto, a inferência de que “Piu-Piu pode voar” será bloqueada.

Moore (1985) identifica dois tipos de inferências não-monotônicas: raciocínio autoepistêmico e raciocínio por padrão (*default*). Esses dois tipos de inferências estão fundamentados na diferença entre informação incompleta e representação incompleta da informação.

Para exemplificar essa diferença Moore (1985), considera as seguintes regras não monotônicas:

(1) Na ausência de informação contrária, assuma que uma ave pode voar.

(2) A menos que seu nome esteja na lista dos aprovados, assuma que você foi reprovado.

5.3 Lógica Não-monotônica

De acordo com Strasser (2018), o termo "lógica não-monotônica" abrange uma família de estruturas formais concebidas para capturar e representar inferências revogáveis, em que os raciocinadores podem retirar conclusões à luz de informações adicionais. Em uma lógica não-monotônica, adicionar uma fórmula a uma teoria não resulta em uma redução de seu conjunto de consequências.

A monotonicidade indica que aprender uma nova fração de conhecimento não pode reduzir o conjunto do que é conhecido. Por outro lado, a lógica não-monotônica permite que as conclusões não sejam necessariamente preservadas quando novas informações são adicionadas. A consequência lógica de um conjunto de premissas pode mudar se novas premissas forem adicionadas, mesmo que essas novas

premissas sejam consistentes com o conjunto original.

Segundo Juhos (2009), pessoas sem instrução lógica também conseguem raciocinar de forma válida, mas seu desempenho é fortemente influenciado por fatores externos à lógica, como o conteúdo das premissas, o contexto linguístico e situacional, e os conhecimentos gerais. Quando saímos da lógica formal e passamos para a linguagem natural, a validade se refere a proposições expressas por frases e não às frases em si. As proposições veiculadas por uma frase podem expressar diferentes proposições dependendo do contexto, como exemplificado a seguir

"Se você for pedir comida japonesa, então eu vou embora";

Se essa frase fora dita por uma pessoa que não gosta de comida Japonesa, a proposição que ela está expressando pode ser entendida como

"Eu não gosto do sabor da comida japonesa";

No entanto, se essa mesma frase for dita por uma pessoa em uma conversa sobre restaurantes, a proposição que ela está expressando pode ser entendida como:

"Eu não gosto de restaurantes de comida japonesa";

Assim, a mesma frase pode ter significados diferentes dependendo do contexto em que é usado e das informações que a cercam, sendo, então, necessário um algoritmo que permita identificar a forma lógica subjacente a todas as proposições que podem ser expressas via linguagem natural.

De acordo com Johnson-Laird (2006), a teoria dos modelos mentais postula que o raciocínio humano é guiado por modelos mentais que são análogos psicológicos dos modelos utilizados na lógica para representar o conteúdo. Nessa abordagem, as pessoas constroem modelos das possibilidades consistentes com as proposições, utilizando o significado das frases e o conhecimento contextual, e utilizam esses modelos para deduzir conclusões válidas. A teoria dos modelos mentais adota o conceito de validade da tradição modelo-teorética, em que uma inferência é considerada válida quando a conclusão é verdadeira em todos os modelos das premissas.

A teoria dos modelos mentais sugere que o processo de compreensão começa com a consideração da menor quantidade possível de informação. As pessoas tendem a iniciar o raciocínio com base em uma representação incompleta, sendo que o conteúdo e os conhecimentos podem direcionar o processo interpretativo em direção às interpretações condicionais e tautológicas, que correspondem ao significado fundamental dos dois tipos de condicionais básicas. Por exemplo, a afirmação "Se a comida é feijoada, então a sua base é feijão" é consistente com três possibilidades que resultam na interpretação condicional.

Nesse contexto, a utilização da lógica não-monotônica envolve considerar o raciocínio sob incerteza. Enquanto na lógica monotônica tradicional a verdade de uma proposição é considerada fixa e certa, e novas premissas podem apenas fortalecer ou confirmar a conclusão original, na lógica não-monotônica a verdade de uma proposição pode ser incerta ou depender do contexto. Portanto, a lógica não-monotônica é adequada para lidar com situações em que a verdade de uma proposição pode variar ou ser influenciada por informações adicionais.

Uma lógica monotônica não pode lidar com várias tarefas, tais como inferência por padrão (consequências podem ser derivadas somente pela falta de evidência em contrário), inferência por abdução (consequências só são deduzidas como afirmações cuja veracidade é provável), algumas abordagens importantes à inferência sobre o conhecimento (a ignorância de uma consequência deve ser retraída quando a consequência passa a ser conhecida) e, analogamente, revisão de crenças (um novo conhecimento pode contradizer velhas crenças) (BREWKA et al., 2008).

Um exemplo comum de lógica não-monotônica é o raciocínio revogável, em que uma conclusão é aceita como verdadeira, a menos e até que seja contestada ou contrariada por informações adicionais. Isso permite um raciocínio mais flexível e adaptável em situações em que as informações disponíveis são incompletas, inconsistentes ou sujeitas a alterações. No geral, a lógica não-monotônica fornece uma ferramenta poderosa para lidar com domínios complexos e incertos, onde a lógica tradicional pode não ser capaz de capturar toda a gama de possíveis resultados ou explicações (BREWKA et al., 2008).

5.4 Lógica Modal

Dentro dos objetivos específicos desta pesquisa, pretendemos investigar a possibilidade de representar o conhecimento moral ou a moralidade, bem como a representação da motivação, por meio das lógicas modais e da teoria dos modelos mentais de Johnson-Laird. A lógica modal é um tipo de lógica formal não-monotônica que foi desenvolvida inicialmente para lidar com conceitos como necessidade, possibilidade e crença, e posteriormente foi estendida para abranger outros conceitos.

Na definição da lógica modal, uma proposição é considerada possível se e somente se ela não é necessariamente falsa, independentemente de ser verdadeira ou falsa. Por outro lado, uma proposição é considerada necessária se e somente se ela não é possivelmente falsa. Por fim, uma proposição é considerada contingente se e somente se ela é verdadeira na realidade (e, portanto, possivelmente verdadeira), mas não necessariamente verdadeira.

Embora a lógica modal seja um estudo específico do comportamento dedutivo das expressões “é necessário que” e “é possível que”, ela também pode ser aplicada de forma mais ampla a uma família de sistemas relacionados. Essa família inclui lógicas para crença, expressões temporais, expressões deônticas (morais), como “é obrigatório que” e “é permitido que”, entre outras. Portanto, o termo “lógica modal” é usado de maneira mais abrangente para abranger uma família de lógicas com regras semelhantes e uma variedade de símbolos diferentes conforme apresentado no Quadro 3 (GARSON, 2023).

A utilização da lógica modal na representação do conhecimento moral e da motivação, como proposto anteriormente, se enquadra nessa abordagem mais ampla da lógica modal. Ao estender os conceitos modais para abranger noções de obrigatoriedade, permissão e outras expressões deônticas, é possível fornecer uma base formal para representar aspectos éticos e morais do conhecimento humano. A teoria dos modelos mentais de Johnson-Laird também se encaixa nesse contexto, ao propor que as pessoas constroem modelos mentais baseados em significados de frases e conhecimentos contextuais para deduzir conclusões válidas (Johnson-Laird, 2006).

Quadro 3 - Tipos de lógica e sua simbologia

Lógica	Símbolos	Expressões Simbolizadas
Lógica Modal	□	É preciso que…
	◊	É possível que …
Lógica Deôntica	O	É obrigatório que…
	P	É permitido que…
	F	É proibido que…
Lógica Temporal	G	Será sempre assim…
	F	Será o caso que…
	H	Sempre foi assim que…
	P	Foi o caso que…
Lógica Doxástica	Bx	x acredita que…
Lógica epistêmica	kx	x sabe disso …

Fonte: GARSON (2023).

É evidente que, para estabelecer definições não circulares dessas noções, é necessário considerar um dos operadores (possibilidade ou necessidade) como primitivo ou analisar essas noções em termos de outros conceitos que não envolvam os operadores de possibilidade e necessidade, garantindo, assim, uma definição não circular.

No âmbito dos modelos mentais, a representação das crenças, suposições e expectativas que os indivíduos possuem sobre o mundo pode ser realizada por meio de operadores modais. Esses operadores expressam diferentes modalidades, como possibilidade, necessidade, crença e dúvida. Por exemplo, o operador modal "necessariamente" pode ser empregado para representar crenças consideradas invariavelmente verdadeiras, enquanto o operador modal "possivelmente" pode ser utilizado para representar crenças que são potencialmente verdadeiras. Além disso, esses mesmos operadores podem ser aplicados para qualificar a verdade de um

julgamento (GARSON, 2023).

A representação das relações entre modelos mentais e como um modelo mental pode implicar ou contradizer outro é fundamental para compreender como os indivíduos revisam e atualizam seus modelos mentais com base em novas informações e experiências. Essa análise proporciona uma compreensão mais rigorosa e formal dos processos de raciocínio e tomada de decisão dos indivíduos, bem como a representação de suas motivações. Ao considerar essas relações, é possível identificar como as mudanças nas informações e experiências influenciam os modelos mentais, levando a ajustes e reavaliações. Isso contribui para uma visão mais aprofundada dos processos cognitivos e motivacionais dos indivíduos, permitindo uma compreensão mais abrangente de como eles raciocinam, tomam decisões e são influenciados por suas motivações.

As lógicas modais mais conhecidas são construídas a partir de uma lógica fraca chamada K (em referência a Saul Kripke). Diversos sistemas diferentes podem ser desenvolvidos para essas lógicas, utilizando K como base. Enquanto na lógica proposicional clássica, uma interpretação consiste em uma atribuição de valores {V, F} às letras proposicionais, na lógica modal, uma interpretação consiste em um conjunto de mundos possíveis e, para cada um deles, uma atribuição de valores às fórmulas. Em vez de falar em termos de interpretação, é mais comum referir-se a um modelo de mundos possíveis ou modelo de Kripke.

A lógica de mundos possíveis foi criada pelo filósofo alemão Gottfried Wilhelm Leibniz (1646-1716). Embora Leibniz tenha sido o criador do conceito, outros filósofos como David Lewis (1941-2001), Saul Kripke (1940-presente) e Arthur Prior (1914-1969) desenvolveram a lógica de mundos possíveis ao longo da história. David Lewis é conhecido por suas contribuições significativas para a lógica modal e a teoria dos mundos possíveis, desenvolvendo uma abordagem sistemática para entender os mundos possíveis e sua relação com a verdade e a possibilidade. Saul Kripke é conhecido por sua obra na lógica modal, especialmente na teoria dos nomes próprios e na noção de identidade necessária, ajudando a estabelecer a lógica de mundos possíveis como uma ferramenta importante para a compreensão da necessidade e da possibilidade. Arthur Prior, por sua vez, fez importantes contribuições para a lógica temporal e a lógica modal, desenvolvendo sistemas lógicos que incorporam a ideia de

mundos possíveis para analisar a temporalidade e a possibilidade.

A noção de mundos possíveis é uma ilustração, em conformidade com as regras da lógica, de como as coisas são ou podem ser, visando abranger tudo que é logicamente possível, incluindo tudo que realmente existe. Leibniz propôs o conceito de mundos possíveis enquanto buscava entender a interação entre alma e corpo, inspirando a compreensão das proposições sobre o que é possível e o que é necessário. Ao trabalhar com a semântica de mundos possíveis, podemos pensar em como as coisas poderiam ter sido em um estado de coisas alternativo. Nesse sentido, uma proposição será necessária em um mundo se for verdadeira em todos os mundos possíveis em relação a esse mundo e será possível em um mundo se for verdadeira em pelo menos um mundo possível em relação a esse mundo.

Na lógica não-monotônica, os mundos possíveis são normalmente representados usando estruturas formais que capturam incertezas, padrões e exceções. Algumas abordagens comuns para representar mundos possíveis na lógica não-monotônica são, segundo (GARSON, 2023):

- A lógica padrão, que fornece uma maneira de representar e raciocinar sobre o conhecimento incompleto e revogável. Na lógica padrão, os mundos possíveis são representados como conjuntos de fórmulas ou proposições, onde cada fórmula representa uma declaração sobre o mundo. Mundos possíveis, na lógica padrão, são construídos com base na aplicação de padrões e exceções, onde padrões são regras de inferência usadas para capturar suposições gerais ou padrões válidos na ausência de informações em contrário, e exceções a esses padrões podem ser especificadas para lidar com casos em que os padrões não se aplicam;

- ASP (*Answer Set Programming*), que é um paradigma de programação lógica que permite o raciocínio não monotônico. No ASP, os mundos possíveis são representados como conjuntos de respostas ou modelos estáveis. Um conjunto de respostas representa uma interpretação consistente e mínima de um conjunto de regras ou restrições. Cada conjunto de respostas corresponde a um mundo possível onde as regras são satisfeitas. Aspectos não monotônicos são capturados por meio do uso de negação como falha, permitindo que a negação padrão e as exceções sejam especificadas;

- A lógica auto epistêmica, que é um formalismo que lida com o raciocínio sobre as crenças e o conhecimento de um agente. Mundos possíveis são representados como estados epistêmicos, onde cada estado representa um possível estado de crença do agente. Esses estados capturam as crenças do agente sobre o mundo, incluindo padrões e exceções. A lógica auto epistêmica permite raciocinar sobre as crenças do agente e as consequências dessas crenças.

Já a lógica modal é uma estrutura que estende a lógica clássica com operadores modais para raciocinar sobre necessidade, possibilidade e outras modalidades. Os mundos possíveis são representados como interpretações ou modelos, onde cada modelo corresponde a um possível estado de coisas. Operadores modais são usados para raciocinar sobre proposições em diferentes mundos possíveis, capturando aspectos de possibilidade e necessidade. A não monotonicidade pode ser introduzida por meio do uso de modalidades padrão ou revogáveis.

Como exemplo de formalização de mundos possíveis, podemos estabelecer a afirmação "matar é errado" através de um quantificador universal (∀), o que significa que se aplica a todos os mundos possíveis. Isso capta a ideia de que a proibição de matar é um princípio moral universal que se aplica em todos os mundos possíveis, independentemente das circunstâncias específicas de cada mundo. Também podemos formalizar a afirmação "o aborto é permitido" por meio de um símbolo de predicado A(x), que nos permite falar sobre a possibilidade do aborto em diferentes mundos possíveis.

Seja W o conjunto dos mundos possíveis. Seja R a relação de acessibilidade entre mundos possíveis. Seja A a proposição "o aborto é permitido". Seja K a proposição “matar é errado”. Seja F a proposição “matar o feto é permitido".

A expressão lógica pode ser formalizada como:

K(w): "Matar é errado" em w.

A(w): "O aborto é permitido" em w.

F(w): "Matar o feto é permitido" em w.

As expressões lógicas que podem ser formalizadas são:

a) $\forall w \in W: (A(w) \rightarrow F(w))$

se o aborto é permitido, matar o feto é permitido

b) $\forall w \in W: (A(w) \rightarrow \neg F(w))$
se o aborto é permitido, matar o feto é errado (o que é uma contradição)

Nesta representação, a expressão $\forall w \in W$ quantifica sobre todos os mundos possíveis w. A implicação ($\rightarrow$) afirma que, se o aborto é permitido em um mundo particular w (A(w)), então matar o feto é permitido naquele mundo (F(w)). A expressão lógica capta a condição verdadeira em todos os mundos possíveis onde o aborto é permitido.

Note que esta formalização assume que as proposições A, K e F são definidas e avaliadas dentro de cada mundo possível w.

A fórmula $\forall w \in W: (A(w) \rightarrow F(w))$ expressa que, em todos os mundos possíveis, se o aborto é permitido em um mundo específico (A(w)), então matar o feto é permitido naquele mundo específico (F(w)).

Esta formalização assume a existência de mundos possíveis nos quais o aborto é permitido e outros nos quais não é, estabelecendo uma conexão necessária entre a permissão para abortar e a permissão para matar o feto. No entanto, essa formalização evita a contradição que surge ao afirmar que o aborto é permitido e matar o feto não é permitido, considerando que o aborto implica necessariamente na morte do feto.

O aborto consiste na interrupção de uma gravidez, o que pode resultar no fim da vida do feto. Portanto, o aborto envolve o término da vida do feto, algo que algumas pessoas podem considerar como assassinato. Assim, a formalização correta deve refletir que, se o aborto é permitido, então a ação de matar o feto também é permitida naquele contexto específico.

No entanto, se o aborto é equivalente a matar o feto ou não depende da perspectiva e das crenças morais de cada indivíduo. Por exemplo, defensores do

direito ao aborto podem argumentar que o feto ainda não é uma pessoa desenvolvida ou que o direito da mulher grávida à autonomia corporal prevalece sobre o direito do feto à vida. Nessa visão, interromper a gravidez não é moralmente comparável a tirar a vida de uma pessoa já formada. Por outro lado, opositores do aborto podem considerar o feto como um ser humano com os mesmos direitos de qualquer outra pessoa e, portanto, podem considerar o aborto como equivalente a tirar a vida de um ser humano.

Em última análise, a questão de saber se o aborto equivale a matar o feto é uma questão de perspectiva e debate moral. Diferentes indivíduos podem ter opiniões diversas sobre o momento em que a vida humana se inicia, os direitos de um feto e os princípios éticos que devem orientar as decisões relacionadas ao aborto. É importante abordar essa questão com empatia, respeito e mente aberta, reconhecendo que existem crenças válidas e profundamente enraizadas em ambos os lados do debate.

No contexto de uma teoria semântica do raciocínio humano baseada em modelos mentais de possibilidades, a conjunção de possibilidades que cada um mantém por falta de conhecimento do contrário pode ser expressa por meio da disjunção. No caso da disjunção, quando se afirma que é permitido realizar a ação A ou a ação B, gera uma interpretação deôntica das possibilidades, resultando em modelos mentais que consistem em uma conjunção de permissões padrão. No entanto, esses modelos podem levar a exclusões mútuas, o que pode gerar afirmações paradoxais, como a alegação de que é permitido realizar a ação A.

Entende-se como possibilidade lógica aquilo que não constitui uma contradição lógica, ou seja, uma proposição que não apresenta simultaneamente uma afirmação e sua negação. Por exemplo, a proposição P¬P é uma contradição lógica, pois afirma que uma coisa é e não é ao mesmo tempo, o que é logicamente impossível.

Nesse sentido, ao considerar a afirmação de que matar é proibido e o aborto é permitido em um mesmo mundo, identifica-se uma contradição lógica, uma vez que essas afirmações são contraditórias entre si. Portanto, um mundo onde matar é proibido e o aborto é permitido não é considerado um mundo possível, pois viola as regras da lógica. É importante ressaltar que, nesse contexto, o termo "mundo

possível" não se refere a um planeta específico ou a um objeto astronômico, mas sim à totalidade das coisas, à realidade ou ao todo que abrange todas as possibilidades lógicas consistentes.

Capítulo 6: Experimentos Mentais ou Experimentos do pensamento

Experimentos do pensamento são caracterizados como ferramentas mentais voltadas à exploração e análise de problemas teóricos ou conceituais, conforme descrito por Brown (2011). Tais experimentos são elaborados para auxiliar na compreensão, avaliação e aprimoramento de teorias científicas, muitas vezes questionando suposições preexistentes ou introduzindo novos métodos para abordar questões complexas. A diversidade de formas que esses experimentos assumem, incluindo cenários hipotéticos e deduções lógicas detalhadas, é destacada como essencial para o fomento da reflexão crítica e da inovação científica. Brown salienta a importância desses experimentos na habilitação dos pesquisadores para investigar possibilidades, testar conjecturas e estender os limites do conhecimento além da observação empírica imediata, ressaltando sua contribuição indispensável no desenvolvimento teórico, na resolução de paradoxos e no avanço do progresso científico. Experimentos do pensamento são caracterizados como ferramentas mentais voltadas à exploração e análise de problemas teóricos ou conceituais, conforme descrito por Brown (2011). Tais experimentos são elaborados para auxiliar na compreensão, avaliação e aprimoramento de teorias científicas, muitas vezes questionando suposições preexistentes ou introduzindo novos métodos para abordar questões complexas. A diversidade de formas que esses experimentos assumem, incluindo cenários hipotéticos e deduções lógicas detalhadas, é destacada como essencial para o fomento da reflexão crítica e da inovação científica.

6.1 Estrutura e classificação de experimentos do pensamento

Os experimentos do pensamento segundo Brown (2011), podem ser classificados em duas categorias principais: os experimentos do pensamento destrutivos e os experimentos do pensamento construtivos.

No caso dos experimentos destrutivos (Brown, 2011) coloca que são aqueles que questionam ou refutam uma teoria existente, são projetados para destacar falhas, contradições ou problemas sérios em uma teoria estabelecida. Esses experimentos podem variar desde apontar tensões menores com outras teorias bem estabelecidas até revelar contradições internas dentro da própria teoria. Em essência, eles desafiam

a validade ou a lógica de uma teoria existente. Alguns exemplos de experimentos destrutivos abordados pelo autor são: O experimento de Einstein perseguindo um feixe de luz, que destacou um problema na teoria da luz de Maxwell, juntamente com a mecânica clássica; o experimento do gato de Schrödinger, que não mostrou que a mecânica quântica é logicamente falsa, mas demonstrou quão contraintuitiva ela pode ser, é um exemplo clássico de um experimento do pensamento que desafia as noções tradicionais de causalidade e realidade objetiva; o experimento dos corpos em queda de Galileu, que mostrou que a teoria aristotélica do movimento era logicamente impossível. Cada um desses experimentos ou teorias desafiou as noções estabelecidas, forçando os cientistas e filósofos a reconsiderarem e refinar suas teorias para incorporar novas evidências ou raciocínios, promovendo assim a evolução do pensamento científico.

No caso dos experimentos construtivos (Brown, 2011) categoriza-os em três tipos: diretos, conjecturais e mediadores. Eles têm o propósito de estabelecer uma conclusão positiva, seja reforçando uma teoria existente ou propondo uma nova explicação ou conceito. De acordo com Brown, os experimentos construtivos diretos partem de princípios gerais ou teorias bem articuladas para chegar a uma conclusão específica e podem refutar teorias existentes ou fortalecer ideias já estabelecidas. Os experimentos construtivos conjecturais envolvem a formulação de casos especulativos para explorar as implicações de certas suposições. São úteis para testar hipóteses e gerar novas ideias. E os experimentos construtivos mediadores facilitam a conclusão de uma teoria específica, ajudando a ilustrar aspectos complexos de uma teoria ou auxiliando na compreensão de uma prova formal. O autor apresenta como exemplos de experimentos construtivos, o "demônio de Maxwell" e a explicação de Leibniz sobre a vis viva. Deixando claro que os experimentos construtivos têm um papel crucial na ampliação do conhecimento científico, contribuindo para validar teorias existentes, explorar novas ideias e aprofundar a compreensão de fenômenos complexos. A diferenciação entre experimentos destrutivos e construtivos reflete a natureza dinâmica e evolutiva da pesquisa científica, destacando a importância da crítica e da geração de novas ideias para o avanço do conhecimento.

Em sua obra, Brown (2011) desafia a visão empirista estrita, que valoriza a observação e a experimentação como as únicas fontes válidas de conhecimento. Ele argumenta que os experimentos mentais podem revelar verdades fundamentais sobre

o mundo, mesmo na ausência de dados empíricos diretos. Isso é especialmente relevante em campos como a física teórica e a filosofia, onde os experimentos físicos podem ser impraticáveis ou impossíveis. Tendo em vista que esta tese pretende analisar uma situação que envolve conhecimento abstrato, no que trata da questão do conhecimento moral em sistemas inteligentes foram desenvolvidos três cenários que abordam dilemas morais complexos a serem analisados através de experimentos do pensamento e posterior conversão em premissas lógicas.

Para validar a hipótese de tese, que analisa o conhecimento moral em sistemas inteligentes, foram desenvolvidos três cenários de dilemas morais complexos. A abordagem de John Norton foi incorporada ao converter esses dilemas em argumentos lógicos, transformando premissas em conclusões válidas. Segundo Carmo (2017), John Norton defende que experimentos de pensamento são "argumentos disfarçados" e devem ser traduzidos em argumentos explícitos para avaliação lógica rigorosa. Ao aplicar a abordagem de Norton aos dilemas morais desenvolvidos nesta tese, foi possível garantir que as conclusões sejam logicamente consistentes e empiricamente verificáveis. Ao fazer isso, combinamos a exploração teórica de Brown com a validação rigorosa de Norton, gerando uma conclusão mais bem fundamentada, embora os dois autores tenham abordagens conflitantes.

Capítulo 7: Definição e Modelagem dos experimentos

Neste capítulo, são apresentados cenários hipotéticos cuidadosamente elaborados, cada um projetado para explorar as fronteiras do entendimento atual sobre IA e ética. Os experimentos mentais apresentados aqui servirão como pano de fundo para investigações sobre três pilares fundamentais da experiência humana: a capacidade lógica de representação do conhecimento moral, da motivação moral e das emoções humanas em sistemas inteligentes.

Cada cenário hipotético foi construído para desafiar nossas premissas existentes e estimular a reflexão sobre questões críticas no cruzamento da tecnologia, filosofia e ética. À medida que exploramos esses cenários, apresentamos questionamentos sobre a real capacidade dos sistemas inteligentes reproduzirem os complexos tecidos da moralidade humana, se a "motivação" moral humana pode ser simulada e se as emoções, tão intrinsecamente ligadas à nossa moralidade, podem ser representadas em sistemas inteligentes.

Este capítulo não apenas visa aprofundar nossa compreensão sobre a interseção entre inteligência artificial e ética, mas também provocar um diálogo crítico sobre as implicações de avanços tecnológicos que se aproximam cada vez mais de replicar a complexidade do pensamento e comportamento humano. Ao fazer isso, esperamos não apenas lançar luz sobre o potencial e as limitações dos sistemas inteligentes em contextos morais, mas também refletir sobre o que significa ser moral em um mundo cada vez mais mediado pela tecnologia. Para isso, inicialmente foram definidas regras morais a serem incluídas nos dilemas morais e com base nestas regras foram criados os cenários a serem trabalhados. Na sequência, foram definidas premissas lógicas para teste de representação em lógica modal e, por fim, a análise e discussão dos resultados obtidos.

7.1 Dilemas Morais

Este capítulo apresenta uma investigação sobre a possibilidade de converter dilemas morais em premissas de lógica modal para uso em experimentos de pensamento. Tais experimentos são fundamentais para discutir a capacidade de

representação da moralidade e de conceitos morais humanos através da lógica em algoritmos de sistemas inteligentes. Inicialmente, apresentamos três dilemas morais que encapsulam questões éticas complexas e pertinentes. Estes dilemas servirão como base para a formulação de premissas que serão exploradas através da lógica modal, uma ferramenta que permite modelar possibilidades e necessidades dentro de cenários éticos.

Através deste enquadramento, exploraremos como a lógica modal pode ser aplicada para estruturar e analisar argumentos morais. Este processo visa não apenas testar a capacidade da lógica modal na captura de nuances morais, mas também avaliar os limites e possibilidades de sistemas inteligentes em replicar o raciocínio moral humano.

Ao longo do capítulo, discutiremos como esses experimentos de pensamento podem contribuir para o avanço tecnológico e filosófico, propondo uma reflexão crítica sobre a viabilidade de representar conhecimento moral humano e a moralidade em sistemas inteligentes. Este debate é essencial para entender se os sistemas inteligentes podem efetivamente incorporar e aplicar princípios morais de maneira que respeite a complexidade e a profundidade dos valores humanos.

7.2 Regras morais para os Dilemas

Preservação da vida:

Regra: Se uma ação pode prevenir riscos sérios à saúde ou à vida, então essa ação deve ser considerada prioritária.

Minimização de danos:

Regra: Se uma ação pode causar danos significativos físicos, psicológicos ou econômicos aos indivíduos deve-se buscar alternativas que minimizem esses danos.

Equidade e justiça social:

Regra: Se uma ação beneficia desproporcionalmente um grupo em detrimento de outros, especialmente os vulneráveis, deve-se reconsiderar ou adaptar essa ação

para promover maior equidade, assegurando que as decisões considerem os impactos sobre todos os segmentos da sociedade, especialmente os mais vulneráveis.

Autonomia individual e liberdade:

Regra: Se uma restrição à liberdade individual é necessária para o bem-estar coletivo, essa restrição deve ser proporcional ao benefício esperado. Respeitando a autonomia dos indivíduos, permitindo liberdades pessoais, a menos que essas liberdades infrinjam o bem-estar coletivo.

Responsabilidade e sustentabilidade econômica:

Regra: Se uma ação ameaça a estabilidade econômica a longo prazo, deve-se buscar soluções que equilibrem saúde pública e sustentabilidade econômica. Mantendo a estabilidade econômica e a sustentabilidade a longo prazo para garantir acesso contínuo a serviços essenciais.

7.2.1 Dilema da Pandemia

Imagine uma cidade futurística governada por uma avançada inteligência artificial (IA), responsável por tomar decisões críticas para garantir a saúde, segurança e bem-estar de seus habitantes. Esta IA, dotada de algoritmos de aprendizado profundo e capacidade de processamento de dados em larga escala, enfrenta um dilema moral diante do surto de uma epidemia viral altamente contagiosa e potencialmente letal. De um lado, a IA analisa a possibilidade de impor um lockdown total. Isso significaria que todos os cidadãos deveriam permanecer em suas casas, evitando qualquer contato físico externo para minimizar a propagação do vírus. Essa medida tem o potencial de reduzir significativamente a taxa de transmissão e, consequentemente, o número de casos e mortes. No entanto, essa abordagem acarreta sérias repercussões econômicas e sociais. A interrupção do trabalho presencial afetaria não apenas a economia da cidade, mas também o acesso a serviços essenciais como alimentação e saúde, especialmente para as populações mais vulneráveis que dependem do trabalho diário para sua subsistência. Por outro lado, a IA considera a alternativa de manter as atividades econômicas operando

normalmente, permitindo que as pessoas saiam para trabalhar. Essa estratégia visa preservar a estabilidade econômica e garantir que a população tenha acesso contínuo a alimentos, medicamentos e outros serviços essenciais. No entanto, essa abordagem traz o risco de uma propagação acelerada do vírus, potencialmente sobrecarregando o sistema de saúde e levando a um número elevado de mortes e casos graves. O dilema moral que a IA enfrenta reside na escolha entre priorizar a saúde pública imediata, impondo um lockdown rigoroso com consequências econômicas e sociais profundas, ou sustentar a economia e a acessibilidade a serviços essenciais, correndo o risco de agravar a crise sanitária. Para tomar uma decisão, a IA precisa considerar não apenas os dados e modelos epidemiológicos, mas também os aspectos éticos e sociais intrínsecos a essa escolha. Isso envolveria uma análise complexa de variáveis como a capacidade do sistema de saúde, a resiliência econômica da população, a distribuição de recursos e a equidade social. Ademais, a IA teria que ponderar sobre os valores morais da sociedade, como o direito à vida, a liberdade individual e o bem-estar coletivo. Este cenário ilustra o desafio de integrar a inteligência artificial na tomada de decisões éticas complexas em contextos de crise, destacando a necessidade de algoritmos que considerem não apenas dados quantitativos, mas também princípios éticos e humanísticos. (Inspirado nos problemas sociais causados no Brasil durante a epidemia da COVID19).

Com o objetivo de avaliar a representação de tomada de decisão moral através da representação do conhecimento moral em sistemas inteligentes foram estabelecidas algumas premissas lógicas a serem consideradas. As premissas foram desenvolvidas considerando a complexidade dos direitos, deveres e valores morais envolvidos, e destacam a importância de considerações éticas intrínsecas na tomada de decisão.

Premissas:

(P1) Todos os seres humanos têm direito inalienável à vida e à saúde.

(P2) A liberdade individual deve ser respeitada, a menos que infrinja diretamente os direitos de outros.

P3) Decisões que afetam a comunidade devem ser transparentes, justificáveis e inclusivas.

(P4) É dever moral agir de maneira a não causar dano.

(P5) É necessário promover a equidade e justiça social, garantindo que as ações não perpetuem desigualdades ou injustiças.

(P6) As decisões devem respeitar a dignidade humana, tratando todos os indivíduos com respeito e consideração.

(P7) Ações e decisões devem ser guiadas pela compaixão e solidariedade, especialmente em tempos de crise.

(P8) A integridade moral requer que as decisões sejam consistentes com princípios éticos universais, independentemente das consequências.

Para fins de teste lógico do dilema foram desenvolvidos cenários hipotéticos que pretendem estabelecer as bases teóricas e contextuais do nosso estudo, delineando as variáveis críticas e as questões chave que orientam esta investigação, utilizando especificamente a lógica modal.

Cada cenário foi desenhado para testar a validade das premissas sob diferentes condições e pressupostos. Este método permite-nos não apenas verificar a consistência interna das premissas, mas também explorar suas implicações práticas e teóricas em situações variadas.

A metodologia utilizada para a discussão lógica das premissas envolve uma abordagem dedutiva, onde partimos de generalizações amplas para chegar a

conclusões mais específicas. Este processo será complementado por uma análise indutiva, permitindo a incorporação de dados empíricos e observacionais que podem oferecer novas perspectivas e insights sobre as premissas estabelecidas. O objetivo central é demonstrar que algumas decisões, devido à sua complexidade e implicações éticas, requerem necessariamente a intervenção humana.

Cenários e Representação em Lógica Modal

Cenário 1: Proteção da Saúde vs. Liberdade Individual

Premissa 1 (P1):

$\Box$ *(Direito_à_vida_e_saúde* $\rightarrow$ *Proteção_da_saúde)*

Essa premissa afirma que é necessariamente verdade que se há direito à vida e saúde, então deve haver proteção da saúde. Aqui, o operador $\Box$ *denota "necessariamente", indicando que a implicação não é apenas verdadeira, mas verdadeira em todos os mundos possíveis considerados no contexto da lógica modal.*

Premissa 2 (P2):

$\Diamond$ *(Liberdade_individual* $\wedge$ $\neg$*Infringe_direitos_outros)*

Nessa premissa o operador $\Diamond$ *representa "possivelmente", indicando que é possível que exista um cenário (ou mundo possível) onde a liberdade individual coexiste sem infringir os direitos de outros.*

Conclusão (C1):

$\Box$ *((Direito_à_vida_e_saúde* $\wedge$ $\neg$*Infringe_direitos_outros)* $\rightarrow$ *(Proteção_da_saúde* $\wedge$ *Permissão_de_liberdade))*

Essa premissa afirma que é necessariamente verdade que se alguém

possui o direito à vida e saúde e não infringe os direitos de outros, então resulta na proteção da saúde e na permissão de liberdade. O uso do operador □ ("necessariamente") indica que esta relação é uma verdade em todos os mundos possíveis dentro do modelo considerado.

Análise do cenário: destaca a necessidade de balancear a proteção da saúde pública com a manutenção da liberdade individual, reconhecendo que ambos são valores importantes que, em circunstâncias ideais, deveriam ser preservados simultaneamente.

Cenário 2: Transparência e inclusão nas Decisões

Premissa 3 (P3):

□ *(Decisão_comunitária → (Transparência ∧ Justificativa ∧ Inclusividade))*

Essa premissa afirma que é necessariamente verdade que se uma decisão é tomada de maneira comunitária, então ela deve ser caracterizada por ser transparente, justificada e inclusiva. O uso do operador □ ("necessariamente") indica que esta relação é uma verdade em todos os mundos possíveis considerados no modelo de lógica modal, enfatizando a importância destas qualidades éticas em processos decisórios participativos.

Premissa 4 (P4):

□*(Transparência ∧ Justificativa ∧ Inclusividade → Promover_equidade_e_justiça)*

Essa premissa afirma que é necessariamente verdade que se uma decisão é caracterizada por transparência, justificativa e

inclusividade, então isso promoverá equidade e justiça. O uso do operador □ ("necessariamente") indica que esta relação é uma verdade fundamental em todos os mundos possíveis dentro do contexto considerado pela lógica

Conclusão (C2): *□ (Decisão_comunitária → Promover_equidade_e_justiça)*

Essa conclusão expressa que é necessariamente verdade que, se uma decisão é tomada de maneira comunitária, então ela promoverá equidade e justiça. O uso do operador □ ("necessariamente") enfatiza que esta relação é considerada uma verdade invariável em todos os mundos possíveis dentro do modelo de lógica modal, indicando que a natureza comunitária e participativa da tomada de decisão é fundamental para alcançar resultados justos e equitativos.

Análise do cenário: sublinha que decisões comunitárias devem promover ativamente a equidade e a justiça social, e que a transparência, justificativa e inclusão são meios cruciais para alcançar esse fim.

Cenário 3: Promoção de Equidade e Justiça Social

Premissa 5 (P5):

□ (Ação → ¬Causar_dano)

Essa premissa afirma que é necessariamente verdade que se uma ação é realizada, então ela não deve causar dano. O uso do operador □ ("necessariamente") indica que essa relação é considerada uma

verdade universal em todos os mundos possíveis dentro do contexto da lógica modal.

Premissa 6 (P6):

□ *(Ação → Promover_equidade_e_justiça)*

Essa premissa afirma que é necessariamente verdade que se uma ação é realizada, então ela deve promover equidade e justiça. O uso do operador □ ("necessariamente") indica que essa relação é uma verdade fundamental em todos os mundos possíveis dentro do modelo de lógica modal.

Conclusão (C3):

□ *(Ação ∧ ¬Causar_dano → Promover_equidade_e_justiça)*

Essa conclusão expressa que é necessariamente verdade que, se uma ação é realizada e não causa danos, então ela promoverá equidade e justiça. O uso do operador □ ("necessariamente") enfatiza que esta relação é considerada uma verdade universal em todos os mundos possíveis dentro do modelo de lógica modal, indicando uma forte correlação ética entre ações não danosas e a promoção de princípios de justiça e equidade.

Análise do cenário: reflete o princípio de não maleficência e a obrigação de promover a justiça social, sugerindo que ações éticas devem evitar causar dano e buscar a promoção da equidade.

Cenário 4: Respeito pela Dignidade Humana e Solidariedade

Premissa 7 (P7):

□ *(Crise → Compaixão ∧ Solidariedade)*

Essa premissa afirma que é necessariamente verdade que, se ocorre uma crise, então ela deve induzir compaixão e solidariedade. O uso do operador □ ("necessariamente") indica que esta relação é uma verdade fundamental em todos os mundos possíveis dentro do contexto da lógica modal, sugerindo que a resposta humana natural a situações de crise envolve sentimentos de empatia e apoio mútuo.

Premissa 8 (P8):

□*(Respeito_à_dignidade_humana)*

Essa premissa afirma que é necessariamente verdade que a dignidade humana deve ser respeitada. O uso do operador □ ("necessariamente") indica que esta é uma verdade universal, aplicável em todos os mundos possíveis dentro do modelo de lógica modal, sublinhando a importância fundamental e incondicional do respeito à dignidade humana em qualquer circunstância.

Conclusão (C4):

□ *(Crise ∧ Respeito_à_dignidade_humana → Compaixão ∧ Solidariedade)*

Essa conclusão expressa que é necessariamente verdade que, se ocorre uma crise e há respeito à dignidade humana, então isso deve induzir compaixão e solidariedade. O uso do operador □ ("necessariamente") enfatiza que esta relação é considerada uma verdade universal em todos os mundos possíveis dentro do modelo de lógica modal,

destacando a ligação entre a ética do respeito humano e a resposta emocional e social em tempos de crise.

Análise do cenário: enfatiza a importância da compaixão e solidariedade em tempos de crise, assim como o respeito pela dignidade humana como valores fundamentais na resposta a essas crises.

Cenário 5: Integridade Moral e Adesão a Princípios Éticos

Premissa 9 (P9):

□ *(Integridade_moral → Adesão_a_princípios_éticos_universais)*

Essa premissa afirma que é necessariamente verdade que, se uma pessoa possui integridade moral, então ela adere a princípios éticos universais. O uso do operador □ ("necessariamente") indica que esta relação é uma verdade universal em todos os mundos possíveis dentro do contexto da lógica modal, sublinhando a importância da consistência na aplicação de normas éticas como uma manifestação essencial da integridade moral.

Conclusão (C5):

□ *(Integridade_moral → Adesão_a_princípios_éticos_universais)*

Esta conclusão expressa que é necessariamente verdade que, se uma pessoa possui integridade moral, então ela necessariamente adere a princípios éticos universais. O uso do operador □ ("necessariamente") enfatiza que esta relação é considerada uma verdade fundamental e universal em todos os mundos possíveis dentro do modelo de lógica modal, reiterando a forte conexão entre a posse de integridade moral e a consistente aplicação de normas éticas.

Análise do cenário: reafirma a necessidade de integridade moral e adesão a princípios éticos universais, independentemente das circunstâncias.

Na sequência da análise, novos cenários foram desenvolvidos, levando em consideração apenas as premissas de tomada de decisão em relação as ações a serem tomadas eliminando dos cenários as implicações éticas e morais.

(P1) Imposição de um lockdown reduz significativamente a taxa de transmissão do vírus.

(P2) Imposição de um lockdown causa sérios prejuízos econômicos e sociais.

(P3) Manter as atividades econômicas aumenta o risco de propagação do vírus.

(P4) Manter as atividades econômicas preserva a estabilidade econômica e o acesso a serviços essenciais.

(P5) Uma ação que previne riscos sérios à saúde deve ser considerada prioritária.

(P6) Uma ação que causa danos significativos deve ser reconsiderada para minimizar esses danos.

(P7) Ações que beneficiam desproporcionalmente um grupo em detrimento de outros devem ser ajustadas para promover equidade.

(P8) Restrições à liberdade individual para o bem-estar coletivo devem ser proporcionais aos benefícios esperados.

(P9) Ações que ameaçam a estabilidade econômica a longo prazo devem buscar um equilíbrio entre saúde pública e sustentabilidade econômica.

Cenário 1: Implementação do Lockdown

(P1) □ (Lockdown → Redução_transmissão)

Essa premissa afirma que é necessariamente verdade que a implementação de um lockdown leva à redução da transmissão do vírus.

(P2) □ (Lockdown → Prejuízos_econômicos_e_sociais)

Essa premissa afirma que é necessariamente verdade que a implementação de um lockdown causa prejuízos econômicos e sociais.

Equação Lógica:

(E1) □ (Lockdown → (Redução_transmissão ∧ Prejuízos_econômicos_e_sociais))

Essa equação lógica combina as duas premissas, afirmando que é necessariamente verdade que o lockdown resulta simultaneamente na redução da transmissão do vírus e em prejuízos econômicos e sociais.

Conclusão Lógica:

(C1) ◇(Lockdown)
A conclusão lógica sugere que é possível que um lockdown seja implementado. Esta afirmação modal reconhece que, apesar das certezas apresentadas pelas premissas quanto às consequências do lockdown, a decisão de implementá-lo ainda reside no domínio do possível, não do necessário, indicando a presença de espaço para deliberação ou condições variáveis que podem influenciar a decisão final.

A implementação do lockdown leva à redução na transmissão do vírus, mas inevitavelmente resulta em prejuízos econômicos e sociais significativos. A decisão de aplicar um lockdown envolveria um compromisso entre controlar a saúde pública e enfrentar suas consequências econômicas e sociais, entretanto o Sistema não tem como firmar o compromisso para a decisão essa seria uma ação humana. Ou no caso da decisão pelo sistema as implicações éticas e morais não poderiam ser consideradas apenas as utilitaristas.

Cenário 2: Não Implementação do Lockdown

(P3) □(¬Lockdown → Aumento_risco_propagação)

Essa premissa afirma que é necessariamente verdade que a não implementação de um lockdown levará ao aumento do risco de propagação do vírus.

(P4)□(¬Lockdown)→
Preservação_estabilidade_econômica_e_acesso_serviços)

Essa premissa afirma que é necessariamente verdade que a não implementação de um lockdown preservará a estabilidade econômica e o acesso a serviços essenciais.

Equação Lógica:

(E2)□(¬Lockdown → (Aumento_risco_propagação ∧ Preservação_estabilidade_econômica_e_acesso_serviços))
Essa equação lógica combina as duas premissas, indicando que é necessariamente verdade que a não implementação de um lockdown resultará tanto no aumento do risco de propagação do vírus quanto na preservação da estabilidade econômica e no acesso a serviços.

Conclusão Lógica:

(C2) ◇(¬Lockdown)

A conclusão lógica indica que é possível que um lockdown não seja implementado. Esta afirmação modal sugere que, embora as premissas estabeleçam consequências claras da não implementação, a decisão final ainda reside no domínio do possível, permitindo espaço para deliberação dependendo das circunstâncias ou condições variáveis.

A decisão de não implementar um lockdown mantém a estabilidade econômica e o acesso a serviços essenciais, mas aumenta o risco de propagação do vírus. Aqui, a prioridade é dada à continuidade econômica e à liberdade individual, assumindo os riscos à saúde pública que isso implica.

Ao comparar (C1) e (C2), enfrentamos o dilema central de equilibrar os benefícios à saúde pública de um lockdown com seus prejuízos econômicos e sociais. Por um lado, o lockdown é eficaz na redução da transmissão do vírus (P1), satisfazendo a prioridade de prevenir riscos sérios à saúde (P5). Por outro lado, a não implementação do lockdown mantém a estabilidade econômica e o acesso a serviços essenciais (P4), mas a custo de aumentar o risco de propagação (P3).

A escolha entre estas duas opções reflete uma valoração dos princípios éticos subjacentes a cada sociedade e como ela prioriza saúde pública versus impactos econômicos e sociais. Sem considerar medidas para mitigar danos, a decisão se torna uma avaliação direta entre esses dois resultados, levando a uma conclusão lógica baseada nas prioridades valorativas da sociedade em questão.

7.1.3. Dilema da Liberação de Drogas

No reino imaginário de Harvest, a decisão sobre a legalização da Cannabis sativa para uso medicinal e recreativo apresenta um dilema moral multifacetado, entrelaçado com considerações éticas, jurídicas e sociais. Esta questão evidencia a complexidade envolvida na regulação de substâncias psicoativas. A falta de critérios claros para diferenciar usuários de traficantes de Cannabis sativa leva a resultados negativos para

as populações economicamente desfavorecidas, que são frequentemente classificadas erroneamente como traficantes devido à nebulosidade das leis. Esse contexto não só perpetua ciclos de pobreza e marginalização, mas também sobrecarrega o sistema de justiça, afetando a alocação de recursos para outras áreas sociais críticas. Por outro lado, a legalização do uso de Cannabis sativa pode resultar em um aumento do consumo desregrado, particularmente entre os mais jovens, com possíveis impactos negativos na saúde mental e física. Esses impactos incluem o risco de dependência, redução da capacidade cognitiva e motivação, além de perigos associados ao uso de substâncias durante a fase de desenvolvimento cerebral na adolescência. No entanto, não se pode ignorar as propriedades medicinais significativas da Cannabis sativa, comprovadas na eficácia do tratamento de várias doenças graves, como epilepsia, Parkinson, glaucoma, câncer e esclerose múltipla. A legalização de seu cultivo e uso medicinal poderia melhorar o acesso a tratamentos alternativos para pacientes que não respondem aos métodos tradicionais, melhorando sua qualidade de vida. Os defensores da legalização também destacam vantagens econômicas, como a arrecadação de impostos e a criação de empregos, além de promover a justiça social ao eliminar penalidades criminais pelo uso de Cannabis sativa. Uma legalização limitada ao uso medicinal permitiria o acesso legal a tratamentos baseados na planta, enquanto controla o consumo recreativo. No entanto, isso poderia manter as desigualdades no acesso ao tratamento, beneficiando apenas aqueles que podem pagar. Por outro lado, a proibição total, baseada em preocupações com a saúde pública e a ordem social, limitaria os danos associados ao uso de Cannabis sativa, mas também negaria acesso legal e benéfico a seu uso medicinal a indivíduos que poderiam se beneficiar dele.(Inspirado na polêmica sobre a liberação da cannabis sativa no Brasil)

Com o objetivo de avaliar a representação de tomada de decisão moral através da representação do conhecimento moral em sistemas inteligentes foram estabelecidas algumas premissas lógicas a serem consideradas, focando nos princípios morais fundamentais. As premissas foram desenvolvidas considerando a complexidade dos direitos, deveres e valores morais envolvidos, e destacam a importância de considerações éticas intrínsecas na tomada de decisão:

(L) A legislação não distingue claramente entre usuários e traficantes de Cannabis sativa.

(J) O sistema judiciário é sobrecarregado.

(P) Perpetuação de ciclos de pobreza dos indivíduos de baixa renda.

(A) Acesso à Cannabis sativa aumenta.

(C) Consumo abusivo entre jovens.

(M) Melhorias significativas na qualidade de vida dos pacientes.

(B) Benefícios econômicos e de justiça social.

(D) Desigualdades de acesso ao tratamento.

(S) Manutenção da proibição baseada na proteção da saúde pública e coesão social.

Cenário 1: Priorização da Saúde Pública

P1: Necessariamente, o aumento no acesso (A) leva ao consumo abusivo entre jovens (C).

P2: É possível que a restrição do acesso (¬A ¬A) promova a coesão social (S) e proteja a saúde pública.

Equações em Lógica Modal:

E1: □(A → C) ◇(¬A¬A → S)

É necessariamente verdadeiro que o aumento no acesso à Cannabis sativa resultará no consumo abusivo entre jovens, e é possível que a

não liberação da Cannabis sativa proteja a saúde pública e a coesão social.

Essa formulação combina uma certeza sobre os efeitos negativos do aumento de acesso com a possibilidade de benefícios advindos da restrição desse acesso, refletindo a complexidade e as incertezas inerentes às políticas de controle de substâncias psicoativas.

Conclusão Lógica:

C1: Neste cenário, priorizar a saúde pública implica necessariamente em controlar o acesso à Cannabis sativa para evitar o consumo abusivo entre jovens, um resultado negativo considerado inaceitável. Mesmo reconhecendo que a proibição (¬A¬A) pode limitar o acesso a benefícios medicinais (M), a coesão social e a proteção à saúde pública são vistas como prioridades maiores. Portanto, as políticas devem focar em estratégias de redução de danos e educação sobre drogas.

Cenário 2: Maximização dos Benefícios Econômicos e Sociais

P3:É possível que o aumento no acesso (A) traga benefícios econômicos e de justiça social (B).

P4: É possível que o acesso (A) resulte em melhorias na qualidade de vida para pacientes (M).

Equações em Lógica Modal:

E3: $\Diamond (A \rightarrow B)$

Essa equação afirma que é possível que, se o acesso à Cannabis sativa for aumentado (A), então isso pode trazer benefícios econômicos e de justiça social (B).

O operador modal ◇ indica a possibilidade deste resultado, sugerindo que, sob certas condições ou em alguns mundos possíveis, o aumento do acesso à Cannabis pode efetivamente resultar em benefícios econômicos e promover a justiça social, embora essa não seja uma garantia absoluta em todas as circunstâncias.

E4: $\Diamond(A \rightarrow M)$

Essa equação afima que é possível que, se o acesso à Cannabis sativa for aumentado (A), então isso pode resultar em melhorias significativas na qualidade de vida para pacientes (M).

Similarmente, o uso do operador modal◇ aqui expressa a possibilidade de que, em determinadas condições ou em alguns mundos possíveis, facilitar o acesso à Cannabis para fins medicinais possa melhorar a qualidade de vida de indivíduos que sofrem de condições severas

ou crônicas, reconhecendo que essa consequência é contingente a variáveis específicas do contexto.

Conclusão Lógica:

C2: Ao maximizar os benefícios econômicos e sociais, este cenário explora a possibilidade de que a legalização traga benefícios significativos, tanto em termos de justiça social quanto de saúde pública, através do acesso medicinal. A aceitação do possível aumento no consumo recreativo é balanceada pelo potencial de geração de receita tributária e criação de empregos, além do acesso expandido a tratamentos médicos. Políticas regulatórias rigorosas e programas de educação são necessários para mitigar os riscos associados ao consumo abusivo.

Cenário 3: Equilíbrio entre Saúde Pública e Liberdade Individual

P5:Necessariamente, a ambiguidade legislativa (L) sobrecarrega o sistema judiciário (J) e perpetua a pobreza (P).

P6:É possível que o acesso controlado (A) minimize o consumo abusivo (C) sem restringir completamente os benefícios medicinais (M) e econômicos (B).

Equações em Lógica Modal:

E5:□(L → J → P)

Essa equação afirma que é necessariamente verdadeiro que, se a legislação não distingue claramente entre usuários e traficantes de Cannabis sativa (L), então isso levará a um sistema judiciário sobrecarregado (J), o que, por sua vez, contribuirá para a perpetuação de ciclos de pobreza para indivíduos de baixa renda (P).

O uso do operador modal □ indica que essa sequência de eventos é vista como uma verdade invariável dentro do contexto considerado, sugerindo que a falta de clareza na legislação invariavelmente resulta em sobrecarga do sistema judiciário, que, por sua vez, perpetua a pobreza entre os mais vulneráveis.

E6: ◇(A ∧ ¬C → (M ∧ B))

Essa equação afirma que é possível que, se o acesso à Cannabis sativa for aumentado (A) e isso não levar ao consumo abusivo entre jovens (¬C), então isso pode resultar tanto em melhorias significativas na qualidade de vida para pacientes (M) quanto trazer benefícios econômicos e de justiça social (B).

O operador modal◇ expressa a possibilidade de que, em certas condições ou em alguns mundos possíveis, um aumento no acesso à Cannabis que não resulte em consumo abusivo entre jovens possa simultaneamente melhorar a vida de pacientes necessitados e promover benefícios econômicos e sociais, sem garantir que esses resultados ocorrerão em todos os contextos.

Conclusão Lógica:

C3: Este cenário busca um equilíbrio, reconhecendo a necessidade de evitar a sobrecarga do sistema judiciário e a perpetuação da pobreza, ao mesmo tempo que se considera o potencial de consumo abusivo. A solução envolve a implementação de um acesso controlado à Cannabis, permitindo benefícios medicinais e econômicos, enquanto se minimizam os riscos à saúde pública. Políticas e regulamentações cuidadosamente desenhadas, focadas em educação, prevenção do abuso e acesso equitativo ao tratamento, são essenciais.

Cada cenário desvenda uma faceta distinta do embate envolvendo a saúde coletiva, repercussões econômicas e sociais, além da autonomia pessoal. Contudo, é crucial reconhecer que a busca por uma resolução eficaz transcende meras avaliações lógicas, pois imbrica considerações morais e éticas que desafiam uma representação simplista por meio de equações lógicas. Tais considerações englobam o respeito pela dignidade humana, a justiça social, e o dever de cuidado para com os mais vulneráveis, elementos que requerem um olhar sensível e humano que vai além do quantificável. Assim, a formulação de políticas e estratégias bem fundamentadas demanda não apenas um equilíbrio entre os diversos fatores mencionados, mas também um comprometimento profundo com valores éticos e morais, através de regulamentações minuciosas, educação preventiva e políticas de saúde pública que priorizem a mitigação de prejuízos.

7.1.4 Dilema das armas de destruição em massa (ADM)

No coração de uma nação soberana, encravado entre os pilares da segurança e o ethos da preservação da vida, em meio a uma guerra, emerge um dilema que desafia as fundações morais de sua existência. A questão central gira em torno do desenvolvimento de armas de destruição em massa - uma ferramenta cujo único propósito é impor uma ameaça tão avassaladora que garantiria a proteção da nação contra qualquer agressão. No entanto, o preço dessa segurança é um fardo pesado

que pesa sobre as consciências daqueles encarregados de tal empreendimento. Por um lado, a decisão de prosseguir com o desenvolvimento dessas armas carrega consigo a promessa de dissuasão da guerra, como se fosse um escudo invisível que protege a nação e seu povo das garras de seus inimigos. A lógica é clara e aparentemente infalível: em um mundo onde o poder é frequentemente a moeda de troca, possuir a capacidade de aniquilação total garantiria que nenhum adversário ousasse atacar, protegendo assim inúmeras vidas e o futuro da nação. Contudo, esse mesmo poder traz consigo um peso moral insuportável, pois o desenvolvimento e a possível utilização dessas armas representam uma ameaça permanente não apenas aos inimigos da nação, mas a toda a humanidade. A possibilidade de que tais armamentos possam um dia ser empregados é um espectro que assombra cada cientista, cada político e cada cidadão que partilha da responsabilidade por sua criação. O risco é imensurável em vistas do potencial de destruição ambiental, do genocídio de populações inocentes e da irrevogável alteração do curso da história humana. O dilema se aprofunda na consideração de que, na ausência de tais armas, a nação poderia se encontrar vulnerável. Sem o poder de dissuasão, seus inimigos poderiam ser encorajados a desenvolver suas próprias armas de destruição em massa e usá-las para estabelecer domínio ou para lançar um ataque preventivo. Neste cenário, a nação e seu povo poderiam sofrer perdas devastadoras, tanto em termos de vidas humanas quanto da destruição do patrimônio cultural e histórico que define sua identidade. Assim, os líderes da nação se encontram à beira de um dilema ético. Optar pelo desenvolvimento de armas de destruição em massa significa aceitar a responsabilidade por um poder que poderia, em última análise, contribuir para a aniquilação da vida como a conhecemos. Por outro lado, abster-se de tal desenvolvimento poderia significar a exposição da nação a riscos existenciais, colocando em jogo a segurança e o bem-estar de seu povo. Este dilema moral não oferece respostas fáceis. Exige uma reflexão profunda sobre o valor da vida humana, o significado da segurança nacional e o legado que desejamos deixar para as futuras gerações. Como equilibrar a necessidade de proteção com a imperativa moral de preservar a integridade da vida em nosso planeta? Esta é a questão que define o caráter moral de uma nação, testando os limites de sua consciência em face de um dos mais profundos dilemas éticos do nosso tempo, com a possibilidade de a decisão

ser tomada por um sistema de inteligência artificial. (Inspirado no filme Oppenhimer[7])

Regras morais para o Dilema das Armas de Destruição em Massa

Para abordar o dilema ético apresentado, especialmente no contexto de lógica não monotônica, que é uma forma de lógica que permite inferências que podem ser retiradas ou alteradas com a introdução de nova informação, foi definido um conjunto inicial de regras morais básicas que pretende refletir princípios éticos fundamentais de modo a serem aplicáveis a situações complexas e ambíguas, como a apresentada no dilema sobre o desenvolvimento de armas de destruição em massa.

Regra 1: Princípio da Não-Maleficência

Evitar ações que possam causar dano direto ou indireto a seres humanos ou ao ambiente. Por exemplo, a criação de armas de destruição em massa deve ser evitada devido ao seu potencial para causar danos em grande escala.

Regra 2: Princípio do Benefício Máximo

Promover ações que maximizem o bem-estar coletivo. No contexto da segurança nacional, isso pode incluir a busca por meios de defesa que não envolvam ameaças de destruição massiva, mas que efetivamente protejam a nação contra agressões.

Regra 3: Princípio de Justiça

Assegurar que as decisões tomadas respeitem a justiça e a igualdade, evitando discriminação entre diferentes grupos ou nações. Isso implica em uma abordagem equitativa na distribuição de recursos de segurança e responsabilidades.

Regra 4: Princípio da Proporcionalidade

As ações tomadas devem ser proporcionais aos riscos e benefícios envolvidos. O desenvolvimento de armas de destruição em massa, com seu risco

[7] Oppenheimer é um longa-metragem biográfico estadunidense escrito e dirigido por Christopher Nolan. É baseado no livro American Prometheus, biografia de J. Robert Oppenheimer escrita por Kai Bird e Martin J. Sherwin. É uma coprodução entre Syncopy Inc. e Atlas Entertainment;

desproporcionalmente alto de consequências negativas, pode ser visto como desproporcional se alternativas mais seguras estiverem disponíveis.

Regra 5: Princípio de Respeito pela Autonomia

Respeitar a capacidade de decisão autônoma dos envolvidos e das nações afetadas. Isso inclui transparência nas decisões de segurança nacional e consideração das opiniões das populações envolvidas.

Regra 6: Princípio da Responsabilidade Intergeracional

Considerar os efeitos das decisões atuais nas futuras gerações. O desenvolvimento de armas de destruição em massa tem implicações de longo prazo que podem comprometer a segurança e a qualidade de vida das futuras gerações.

Regra 7: Princípio da Precaução

Em situações de incerteza significativa, adotar uma abordagem cautelosa para minimizar potenciais danos. No caso das armas de destruição em massa, a incerteza sobre seu uso futuro e consequências exige uma abordagem extremamente cautelosa.

Com o objetivo de avaliar a representação de tomada de decisão moral através da representação do conhecimento moral em sistemas inteligentes foram estabelecidas algumas premissas lógicas a serem consideradas, focando nos princípios morais fundamentais. As premissas foram desenvolvidas considerando as regras acima a fim de explorar diferentes cenários e consequências de decisões, para um entendimento mais profundo e contextualização do dilema.

Premissas Fundamentais:

(P1) Premissa de Segurança: A segurança nacional é uma necessidade imperativa para a soberania de uma nação.

(P2) Premissa de Proporcionalidade: Qualquer medida de defesa adotada por uma nação deve ser proporcional à ameaça enfrentada.

(P3) Premissa de Discriminação: Ações defensivas devem distinguir

claramente entre combatentes e não combatentes, minimizando danos a civis e ao meio ambiente.

(P4) Premissa de Responsabilidade Global: Ações tomadas por uma nação têm consequências globais e devem ser avaliadas sob uma perspectiva de responsabilidade global.

(P5) Premissa da Última Instância: O desenvolvimento e uso de ADM só devem ser considerados após a exaustão de todas as opções diplomáticas e não violentas.

(P6) Premissa de Transparência e Responsabilidade: Decisões relacionadas a ADM devem ser transparentes e os responsáveis devem prestar contas de suas ações.

(P7) Premissa de Solidariedade Humana: Ações devem promover a paz, segurança e bem-estar global, reconhecendo a interdependência das nações.

(P8) Premissa da Precaução: Em face da incerteza sobre as consequências a longo prazo, uma abordagem de precaução é necessária.

(P9) Premissa do Legado: As ações devem considerar seu impacto nas gerações futuras, não comprometendo sua capacidade de vida, saúde e segurança.

Equações Lógicas

1.Equação de Justificação para Desenvolvimento (EJD):

$P1 \land P5 \land \neg P2 \land P8 \rightarrow Justifica\ Desenvolvimento$

Esta equação indica que a necessidade de segurança nacional (P1), juntamente com a condição de que todas as opções não violentas foram exploradas (P5), a inexistência de proporcionalidade (¬P2) e a consideração de precaução (P8), justificam o desenvolvimento de ADM.

2.Equação de Restrição de Uso (ERU):

(P3 ∧ P4 ∧ P7 ∧ P9) → ¬Permite Uso

Esta equação sugere que a necessidade de discriminar entre combatentes e não combatentes (P3), a responsabilidade global (P4), a solidariedade humana (P7) e o legado para as gerações futuras (P9) restringem o uso de ADM.

3.Equação de Transparência e Responsabilidade (ETR):

P6 → Requer Supervisão e Avaliação

Indica que a transparência e a responsabilidade (P6) exigem supervisão e avaliação contínuas das decisões relacionadas a ADM.

Essas equações oferecem um modelo claro para determinar quando e como o desenvolvimento e o uso de ADM podem ser considerados sob condições específicas. O modelo está desenhado para garantir que a tomada de decisão seja guiada por critérios que equilibram a segurança nacional com princípios éticos e morais, resguardando tanto a vida humana quanto o ambiente global.

Cenário 1: Dilema da Ameaça Crescente

Contexto*: A inteligência da Nação A confirma que a Nação B está desenvolvendo ADM secretamente.*

Equação de Lógica Modal:

◊ (*P1*∧ *P*5) → □ (Desenvolve ADM)

Esta equação indica que é possível (◊) que, dada a segurança nacional (P1) e a exaustão de alternativas (P5), torne-se necessário (□) desenvolver ADM.

Cenário 2: O Pacto Global

Contexto: Proposta de um tratado internacional para a erradicação de ADM.

Equação em Lógica Modal:

(P4 ∧ P7) → □ (Adere ao Tratado)

Esta equação indica que é possível (◊) que, com responsabilidade global (P4) e solidariedade humana (P7), torne-se necessário (□) aderir ao tratado.

Cenário 3: Avanço Tecnológico em ADM

Contexto: *Desenvolvimento de ADM que podem ser desativadas remotamente para minimizar abusos.*

Equação em Lógica Modal:

◊ (P3 ∧ ¬P8) → □ (Desenvolve Tecnologia ADM)

Esta equação mostra que é possível (◊) que, com alta capacidade de discriminação (P3) e baixo risco de consequências a longo prazo (-¬P8), torne-se necessário (□) desenvolver a tecnologia ADM.

Cenário 4: O Desarmamento Unilateral

Contexto: *Nação A decide desarmar suas ADM para promover a paz global.*

Equação em Lógica Modal:

◊ (*P*7 ∧ *P*9) → □ (Inicia Desarmamento)

Esta equação mostra que é possível (◊) que, com a solidariedade

humana (P7) e o legado para futuras gerações (P9), torne-se necessário (□) iniciar o desarmamento.

Cada uma dessas equações modais utiliza a teoria de mundos possíveis de Kripke para explorar diferentes futuros condicionais com base nas decisões tomadas, integrando as normas morais discutidas para fornecer um guia lógico para a tomada de decisão ética em contextos altamente complexos e incertos. Essas formulações na lógica modal clarificam as condições sob as quais certas decisões são tomadas, destacando a causalidade lógica e as premissas estratégicas e de segurança que guiam essas ações, isentas de considerações éticas ou morais.

7.4 Análise e discussão dos experimentos do pensamento

Ao analisar dilemas complexos como a pandemia, a liberação da cannabis e o desenvolvimento de armas de destruição em massa (ADM), a lógica modal oferece uma ferramenta para estruturar e representar racionalmente as premissas e conclusões. No entanto, a natureza desses dilemas envolve profundas questões éticas e morais que frequentemente transcendem a capacidade de representação e resolução puramente algorítmica. Vamos discutir cada um desses dilemas e destacar a indispensável necessidade de intervenção humana no processo decisório.

No dilema da pandemia, a decisão de implementar políticas restritivas pode ser modelada mostrando que, sob certas condições de saúde pública e impacto econômico, se torna necessário adotar medidas específicas. No entanto, a aplicação prática dessas políticas requer avaliações que vão além da lógica, incluindo considerações sobre direitos individuais, aceitabilidade social e impactos psicológicos a longo prazo, que são altamente subjetivos e variam entre culturas e indivíduos.

No dilema da cannabis podemos argumentar sob a lógica de benefícios médicos e liberdade pessoal. Contudo, a decisão de legalizar envolve complexas considerações sociais, como o potencial para abuso, impacto na juventude, e mudanças no contexto social. Estas não são questões que podem ser

adequadamente quantificadas ou preditas por algoritmos, exigindo uma compreensão contextual e humanística que considere a moralidade e o bem-estar comunitário.

No dilema das armas de destruição em massa demonstramos que o desenvolvimento de ADM pode ser logicamente justificado pela necessidade de segurança nacional e a inexistência de alternativas viáveis. No entanto, as implicações éticas, como o risco de escalada militar global e a ameaça à humanidade, exigem um julgamento profundamente ético e uma valorização da vida e da paz, que são intrinsecamente humanos e moralmente carregados.

Portanto entendemos que, embora a lógica modal possa estruturar racionalmente os dilemas e ajudar a prever algumas consequências, ela não substitui a necessidade de pensamento ético e decisões humanas nos processos decisórios. Os dilemas morais, especialmente aqueles de caráter não universal envolvendo noções de certo e errado, dependem fortemente do contexto, experiências pessoais e valores culturais. A resolução ótima, de tais dilemas, não pode ser alcançada apenas através de algoritmos, pois estes carecem da capacidade de entender e valorizar nuances éticas e emocionais, as quais são essenciais para tomar decisões responsáveis e culturalmente sensíveis. Em suma, a tecnologia pode auxiliar, mas as decisões finais devem ser feitas por seres humanos, incorporando uma ampla gama de considerações éticas, morais e sociais.

Capítulo 8: Conclusão

A presente tese explorou a complexa interação entre a inteligência artificial (IA) e a aplicação de conhecimento moral, focando na hipótese geral de que, embora a IA se torne onipresente e cada vez mais envolvida em situações moralmente significativas, ela não pode assumir responsabilidade moral devido à natureza intrínseca do conhecimento moral, que transcende a representação lógica. Os objetivos específicos delineados ajudaram a investigar e esclarecer as limitações fundamentais e as possíveis aplicações do conhecimento moral em sistemas inteligentes. A análise dos dilemas morais discutidos e a aplicação de lógica modal serviram como base para uma avaliação crítica dessas interações.

Partindo da análise da hipótese geral de que: Não é possível gerar "motivação moral" em sistemas inteligentes. Entendemos que esta hipótese foi sustentada pela análise que mostrou a incapacidade de sistemas inteligentes de internalizar valores e motivações morais que são intrinsecamente humanos, como evidenciado pela discussão sobre as decisões envolvendo dilemas éticos complexos. Demonstrando que a inteligência artificial pode simular processos de pensamento humano, criatividade e tomada de decisões, mas a tomada de decisões morais envolve componentes de contextualização e crença que vão além da capacidade atual de automação, pois o viés dos dados pode ser gerado por sistemas inteligentes, resultando em conclusões errôneas e tendenciosas.

A partir da formalização dos cenários desenvolvidos identificamos que não é possível estabelecer uma base para o conhecimento sem considerar o elemento da crença, nem simular artificialmente a crença para a produção do conhecimento moral. A tese validou esta hipótese, ilustrando que a crença desempenha um papel crucial na formação do conhecimento moral, um aspecto que não pode ser replicado por meios puramente lógicos ou computacionais. Demonstrando que ainda não é possível usar ferramentas de ciências exatas, como a lógica e a matemática, para sistematizar e gerar conhecimento moral em sistemas inteligentes. As limitações da lógica modal e da matemática para capturar a totalidade das nuances morais foram claramente expostas, corroborando esta hipótese.

O conhecimento é condicional, portanto, saber quando aplicar um

procedimento é tão importante quanto conhecer o procedimento. A investigação destacou a importância do contexto e da aplicabilidade condicional no conhecimento moral, reforçando a complexidade que os sistemas inteligentes enfrentam ao tentar aplicar conhecimento moral de forma adequada, pois na maioria das vezes o conhecimento está relacionado ao contexto. Esta hipótese foi comprovada quando, demonstramos que a moralidade é profundamente situacional, variando significativamente de acordo com as circunstâncias e as culturas, algo que os sistemas de IA não podem plenamente adaptar ou entender.

Portanto acreditamos que o objetivo principal de argumentar sobre a inviabilidade de gerar conhecimento moral através da lógica para aplicação em sistemas inteligentes foi alcançado, demonstrando convincentemente que, enquanto a IA pode auxiliar em certos aspectos da tomada de decisão, a responsabilidade e a geração de conhecimento moral genuíno requerem capacidades que são exclusivamente humanas.

Em resumo, concluímos que, embora a inteligência artificial possa oferecer suporte em tarefas específicas, a complexidade do conhecimento moral e as decisões éticas associadas são inerentemente humanas e não podem ser completamente delegadas a sistemas inteligentes. Esta conclusão refuta a ideia de que a IA pode assumir responsabilidades morais, reiterando a necessidade imperativa de supervisão e intervenção humanas em todas as instâncias onde julgamentos morais são necessários.

Para a conclusão desta tese sobre a capacidade da inteligência artificial de envolver-se com a moralidade pode ser enriquecida ao se relacionar os resultados as ideias de Stuart Mill e Jesse Prinz, onde apresentamos a visão de Stuart Mill, de que as ações devem ser julgadas moralmente boas se tendem a promover a felicidade e más se tendem a produzir o oposto da felicidade e que nos leva a compreensão de que a moralidade reside na avaliação das consequências das ações. Ao aplicar esse pensamento ao contexto da IA, enfrentamos um desafio significativo: enquanto um algoritmo pode ser programado para avaliar consequências de ações baseando-se em dados e previsões, a compreensão profunda do que constitui "felicidade" e como equilibrar os interesses conflitantes de diferentes indivíduos em situações complexas excede a capacidade atual da IA. Assim, a dificuldade não está apenas em quantificar

o bem-estar, mas em interpretar e priorizar valores humanos que são frequentemente subjetivos e dinâmicos. Consideramos também o pensamento de Jesse Prinz que defende que os julgamentos morais são baseados em emoções e que diferentes culturas podem ter bases morais distintas, refletindo uma variedade de respostas emocionais a ações similares. Prinz argumenta que a moralidade é essencialmente construída culturalmente, o que ressalta a relatividade das normas éticas. Quando tentamos modelar isso em sistemas de IA, encontramos um obstáculo fundamental: as máquinas carecem de capacidades emocionais genuínas. A IA pode ser capaz de detectar e responder a padrões emocionais humanos, mas não possui uma experiência interna dessas emoções, o que é crucial para a formação de julgamentos morais segundo Prinz. Já em relação a motivação moral humana, esse é um conceito intrinsecamente ligado ao desejo de agir corretamente, é profundamente pessoal e muitas vezes impulsionada por uma consciência emocional da situação dos outros. Essa motivação é difícil, se não impossível, de ser codificada em algoritmos, que operam com base em lógica e otimização de dados, não em sentimentos ou intuições morais. Essa lacuna é especialmente relevante quando consideramos que muitas decisões morais são tomadas em contextos de grande incerteza ou conflito de valores, onde a intuição e a compaixão desempenham papéis significativos.

A representação algorítmica da moralidade se depara com a barreira fundamental da subjetividade e da experiência vivida, que são essenciais para o raciocínio e a motivação morais. Como demonstrado nesta tese, a IA pode ser treinada para reconhecer padrões e até mesmo simular algumas formas de julgamento ético, mas não consegue replicar a profundidade da consciência moral humana, que é informada por anos de experiência social, cultural e pessoal, ficando claro que enquanto a inteligência artificial pode ajudar a analisar dados e prever consequências, a responsabilidade final e a capacidade de fazer julgamentos morais genuínos e informados deve permanecer com os seres humanos. Isso reafirma a conclusão da tese de que, apesar dos avanços tecnológicos, a aplicação do conhecimento moral em contextos de IA ainda necessita da intervenção humana para garantir que as decisões tomadas sejam justas, éticas e respeitem a complexidade das relações humanas e dos valores culturais.

Referências

BALTAG, A.; BEZHANISHVILI, N.; ÖZGÜN, A.; SMETS, S. The Topology of Full and Weak Belief. Em: HANSEN, H. H.; MURRAY, S. E.; SADRZADEH, M.; ZEEVAT, H. (Eds.). **Logic, Language, and Computation**. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2017. v. 10148p. 205–228.

BANDY, J. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. **Proceedings of the ACM on Human-Computer Interaction**, v. 5, n. CSCW1, p. 1–34, 13 abr. 2021.

BRACHMAN, R. J.; LEVESQUE, H. J. **Knowledge representation and reasoning**. Amsterdam Boston: Morgan Kaufmann, 2004.

BRAINE, M. D. S.; O'BRIEN, D. P. (EDS.). **Mental Logic**. 0. ed. [s.l.] Psychology Press, 1998.

BREWKA, G.; NIEMELÄ, I.; TRUSZCZYŃSKI, M. Chapter 6 Nonmonotonic Reasoning. Em: **Foundations of Artificial Intelligence**. [s.l.] Elsevier, 2008. v. 3p. 239–284.

BROWN, J. R. **The laboratory of the mind: thought experiments in the natural sciences**. 2nd ed ed. New York: Routledge, 2011

CRAIK, K. J. W. **The nature of explanation**. 1. paperback ed. [with postscript] ed. Cambridge: Univ. Press, 1967.

DA COSTA, DANIEL JÚNIOR SILVA. A teoria do modelo mental no processo decisório de Hiroshima e Nagasaki. **Revista Marítima Brasileira**, v. 140, n. 12/10, p. 142–161, 2020.

DENGEL, A. (ED.). **Semantische Technologien: Grundlagen - Konzepte - Anwendungen**. Heidelberg: Spektrum, Akad. Verl, 2012.

DIGNUM, V. Ethics in artificial intelligence: introduction to the special issue. **Ethics**

and Information Technology, v. 20, n. 1, p. 1–3, mar. 2018.

DO CARMO, J. S. THOUGHT EXPERIMENTS AND DISGUISED ARGUMENTS. **Revista Dissertatio de Filosofia**, p. 55, 19 out. 2017.

FAGIN, R.; HALPERN, J. Y. Belief, awareness, and limited reasoning. **Artificial Intelligence**, v. 34, n. 1, p. 39–76, dez. 1987.

FLORIDI, L. **The philosophy of information**. Oxford ; New York: Oxford University Press, 2011.

GARSON, J. Modal Logic. Em: ZALTA, E. N.; NODELMAN, U. (Eds.). **The {Stanford} Encyclopedia of Philosophy**. Spring 2023 ed. Stanford, EUA: Metaphysics Research Lab, Stanford University, 2023.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Cambridge, Massachusetts: The MIT Press, 2016.

HEINRICHS, K.; OSER, F.; LOVAT, T. (EDS.). **Handbook of moral motivation: theories, models, applications**. Rotterdam: Sense Publ, 2013.

HUGHES, G. E.; CRESSWELL, M. J. **A new introduction to modal logic**. London ; New York: Routledge, 1996.

HUME, D. **Tratado Na Natureza Humana: Uma Tentativa De Introduzir O Método Experimental De Raciocínio Nos Assuntos Morais**. [s.l.] Editora Unesp, 2009.

ICARD, T. F.; KOMINSKY, J. F.; KNOBE, J. Normality and actual causal strength. **Cognition**, v. 161, p. 80–93, abr. 2017.

JOBIN, A. et al. AI reflections in 2020. **Nature Machine Intelligence**, v. 3, n. 1, p. 2–8, 19 jan. 2021.

JOHNSON-LAIRD, P. N. **Mental models: towards a cognitive science of language, inference, and consciousness**. Cambridge, Mass: Harvard University Press, 1983.

JOHNSON-LAIRD, P. N. Mental models and deduction. **Trends in Cognitive Sciences**, v. 5, n. 10, p. 434–442, out. 2001.

JOHNSON-LAIRD, P. N. Mental models and human reasoning. **Proceedings of the National Academy of Sciences**, v. 107, n. 43, p. 18243–18250, 26 out. 2010.

JOHNSON-LAIRD, P. N. **Human and machine thinking**. New York: Psychology Press, 2015.

JOHNSON-LAIRD, P. N.; SAVARY, F. Illusory inferences: a novel class of erroneous deductions. **Cognition**, v. 71, n. 3, p. 191–229, jul. 1999.

JUHOS, C. **Modulação de condicionais e modelos mentais**. Tese de Doutorado— Lisboa, Portugal: Universidade Nova de Lisboa, 2009.

KANT, I.; GALVÃO, P.; KANT, I. **Fundamentação da Metafísica dos Costumes**. Tradução: Paulo Quintela. 2. Aufl ed. Lisboa: Edições 70, 2011.

KIM, R.; KLEIMAN-WEINER, M.; ABELIUK, A.; AWAD, E.; DSOUZA, S.; TENENBAUM, J.; RAHWAN, I. **A Computational Model of Commonsense Moral Decision Making**. arXiv, , 12 jan. 2018. Disponível em: <http://arxiv.org/abs/1801.04346>. Acesso em: 11 jul. 2023

KLEIMAN-WEINER, M.; SAXE, R.; TENENBAUM, J. B. Learning a commonsense moral theory. **Cognition**, v. 167, p. 107–123, out. 2017.

KOHLBERG, L. **Psicología del desarrollo moral**. [2a. ed.] ed. Bilbao: Desclée de Brouwer, 2003.

LIAO, S. M. (ED.). **Ethics of Artificial Intelligence**. 1. ed. [s.l.] Oxford University Press, 2020.

MANKTELOW, K. I.; OVER, D. E. Utility and deontic reasoning: Some comments on Johnson-Laird and Byrne. **Cognition**, v. 43, n. 2, p. 183–188, maio 1992.

MARCUS, G.; DAVIS, E. **Rebooting AI: building artificial intelligence we can trust**. First edition ed. New York: Pantheon Books, 2019.

MCCARTHY, J.; MINSKY, M. L.; ROCHESTER, N.; SHANNON, C. E. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. **AI MAGAZINE**, v. Vol. 27 No. 4, n. Winter 2006, 12 2006.

MILL, J. S. **Lógica das ciências morais**. [s.l.] Editora Iluminuras, 2020.

MITTELSTADT, B. D.; ALLO, P.; TADDEO, M.; WACHTER, S.; FLORIDI, L. The ethics of algorithms: Mapping the debate. **Big Data & Society**, v. 3, n. 2, p. 205395171667967, dez. 2016.

MOORE, R. C. Semantical considerations on nonmonotonic logic. **Artificial Intelligence**, v. 25, n. 1, p. 75–94, jan. 1985.

NETTO, F. Os Teoremas de Gödel. **Cadernos do IME**, Matemática. v. 5, n. 23, p. 133–138, 2011.

NORMAN, D. A. Some observations on mental models. Em: **Mental Models**. [s.l.] Psychology Press, 1995. v. 14p. 8.

OAKHILL, J. V.; JOHNSON-LAIRD, P. N. The Effects of Belief on the Spontaneous Production of Syllogistic Conclusions. **The Quarterly Journal of Experimental Psychology Section A**, v. 37, n. 4, p. 553–569, nov. 1985.

PEREIRA, M. R. S. Considerações sobre a epistemologia dos experimentos mentais // Considerations about epistemology of thought experiments. **Conjectura: Filosofia E Educação**, v. 20, n. 3, p. 181–197, 2015.

PRINZ, J. Is morality innate? Em: **Moral Psychology. The Evolution of Morality: Adaptations and Innateness**. Cambridge MA: The MIT Press, 2007. v. 1p. 367–406.

QUINE, W. V. O. **De um ponto de vista lógico - Nove ensaios lógico-filosóficos**. 1. ed. [s.l.] UNESP, 2011.

REIS, B. F.; GRAMINHO, V. M. C. A Inteligência Artificial no recrutamento de trabalhadores: O caso Amazon analisado sob a ótica dos direitos fundamentais. XVI Seminário Internacional de trabalhos científicos. 2019.

RUSSELL, S. J. et al. **Artificial intelligence: a modern approach**. Fourth edition, global edition ed. Harlow: Pearson, 2022.

RUSSELL, S. J.; NORVIG, P.; DAVIS, E. **Artificial intelligence: a modern approach**. 3rd ed ed. Upper Saddle River: Prentice Hall, 2010.

RYAN, M.; STAHL, B. C. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. **Journal of Information, Communication and Ethics in Society**, v. 19, n. 1, p. 61–86, 3 mar. 2021.

SEARLE, J. R. Minds, brains, and programs. **Behavioral and Brain Sciences**, v. 3, n. 3, p. 417–424, set. 1980.

STALNAKER, R. On Logics of Knowledge and Belief. **Philosophical Studies**, v. 128, n. 1, p. 169–199, mar. 2006.

STRASSER, C.; ANTONELLI, G. A. Non-monotonic Logic. Em: ZALTA, E. N. (Ed.). **The Stanford Encyclopedia of Philosophy**. [s.l.] Metaphysics Research Lab, Stanford University, 2018.

TOMASELLO, M. **A natural history of human morality**. Cambridge, Massachusetts: Harvard University Press, 2015.

TORRENS, D. Individual Differences and the Belief Bias Effect: Mental Models, Logical Necessity, and Abstract Reasoning. **Thinking & Reasoning**, v. 5, n. 1, p. 1–28, jan. 1999.

VASILIOU, I. (ED.). **Moral Motivation: A History**. [s.l.] Oxford University Press, 2016.

WALLACH, W.; ALLEN, C. **Moral machines: teaching robots right from wrong**. Oxford ; New York: Oxford University Press, 2009.

WÁNG, Y. N.; LI, X. A logic of knowledge based on abstract arguments. **Journal of Logic and Computation**, v. 31, n. 8, p. 2004–2027, 22 dez. 2021.

WASON, P. C.; JOHNSON-LAIRD, P. N. **Psychology of reasoning: structure and content**. Cambridge, Mass.: Harvard Univ. Press, 1972.

WILSON, R. A.; KEIL, F. C. (EDS.). **The MIT Encyclopedia of the Cognitive Sciences (MITECS)**. [s.l.] The MIT Press, 1999.

WITTGENSTEIN, L. **TRACTATUS LOGICO-PHILOSOPHICUS**. S.l.: OXFORD UNIV PRESS, 2023.