DESAFIOS DA REPRESENTAÇÃO DO CONHECIMENTO MORAL E DA MOTIVAÇÃO MORAL EM SISTEMAS INTELIGENTES

FLÁVIA BRAGA DE AZAMBUJA[1]; JULIANO SANTOS DO CARMO [2]

[1] Universidade Federal de Pelotas – flaviaazambuja@gmail.com
[2] Universidade Federal de Pelotas – juliano.ufpel@gmail.com

1. INTRODUÇÃO

Com o avanço tecnológico dos sistemas digitais, houve uma evolução dos sistemas computacionais que supostamente "aprendem", o que levanta a necessidade de discutir questões éticas e morais. No entanto, ao relacionar ética e moral com questões tecnológicas, não podemos deixar de reconhecer que as tecnologias não são entidades autônomas, e os dados alimentados nos sistemas são resultado de especificações humanas, que podem conter vieses da natureza humana.

Um fator importante a ser considerado é que os sistemas de aprendizado de máquina podem reproduzir e ampliar os preconceitos daqueles que fornecem conjuntos de dados de treinamento. Além disso, os sistemas podem ser eficientes em relações estatísticas, devido à definição matemática e às ciências exatas envolvidas, mas ineficientes em relações causais. Confundir esses dois conceitos é uma importante fonte de viés. Mesmo que os conjuntos relevantes de normas que representam a moralidade sejam identificados, ainda não está claro como essas normas devem ser representadas.

Considerando a influência humana na configuração dos sistemas e o fato de que já existem sistemas tomando decisões com implicações morais, é crucial que a comunidade acadêmica avance no estudo da ética no desenvolvimento de novas tecnologias. Urge a proposição de modelos que possibilitem o desenvolvimento de linguagens adequadas para transmitir o pensamento e a compreensão da moral aos sistemas autônomos.

Partimos do pressuposto que a sociedade ainda não tem padrões universais ou diretrizes específicas para incorporar normas humanas ou valores morais em sistemas inteligentes autônomos. Segundo TOMASELLO, (2015), a versão exclusivamente humana de cooperação, conhecida como moralidade, aparece na natureza em duas formas análogas. Por um lado, um indivíduo pode se sacrificar para ajudar outro com base em motivos altruístas, como compaixão, preocupação e benevolência; por outro lado, os indivíduos que interagem podem buscar uma maneira equilibrada de beneficiar a todos, baseando-se em motivos imparciais como equidade e justiça. No entanto, o problema fundamental é que, em situações que exigem justiça, geralmente ocorre uma interação complexa entre os motivos cooperativos e competitivos de vários indivíduos. Tentar ser justo significa buscar um equilíbrio entre esses motivos, existem várias maneiras possíveis de fazer isso com base em diferentes critérios.

Portanto, ao programar sistemas autônomos que tomam decisões e podem interagir com o ambiente, é essencial que eles sejam projetados para adotar, aprender e seguir as normas e valores morais da comunidade em que estão inseridos. As ações desses sistemas devem ser transparentes e confiáveis, considerando as situações em que eles operam e os seres humanos que os utilizam. O grande desafio é encontrar mecanismos adicionais para alinhar os

sistemas de inteligência artificial às normas e valores morais humanos, pois eles já estão tomando decisões com implicações morais.

2. METODOLOGIA

Tratando-se de um estudo analítico, com o objetivo de explorar estudos existentes sobre ética em inteligência artificial e sistemas autônomos, foi desenvolvida uma pesquisa bibliográfica através do levantamento ou revisão de obras publicadas sobre a teoria a fim de direcionar e apoiar o trabalho tese que está sendo desenvolvido pelo autor.

3. RESULTADOS E DISCUSSÃO

O estudo da inviabilidade de produção de conhecimento moral através de algoritmos lógicos se justifica diante as seguintes questões: Podemos construir agentes morais artificiais usando aprendizado de máquina? Existe a possibilidade de treinar a Inteligência Artificial (IA) para identificar o bem e o mal, o certo e o errado e depois usá-la para nos ensinar moralidade?

Partindo do princípio de que sistemas não têm viés, as correlações entre os dados podem facilmente transformar sistemas autônomos em sistemas racistas, sexistas e classistas, mesmo quando alimentados com dados imparciais. Isso ocorre porque o preconceito pode ser gerado se um sistema autônomo que "aprende" através da inteligência artificial for alimentado com dados específicos de um grupo. É importante ressaltar que os sistemas de inteligência artificial e aprendizado atuais não conseguem distinguir entre relações estatísticas causais e gerais, nem estabelecer raciocínio causal, não-monotônico, social, moral ou de senso comum. Portanto, esses sistemas podem chegar a conclusões errôneas e tendenciosas. Para resolver essa questão, seriam necessárias representações de conhecimento simbólico explícito, embora ainda não esteja estabelecido como isso pode ser alcançado de forma precisa.

Devido à importância desse tema, recentemente, o Institute of Electrical and Electronics Engineers (IEEE) lançou a discussão intitulada "Iniciativa Global para Considerações Éticas em Inteligência Artificial e Sistemas Autônomos". Essa iniciativa busca a contribuição da área de Ciências Humanas para identificar e estabelecer um amplo consenso sobre questões éticas e recomendações relacionadas a essas tecnologias. De acordo com o documento da IEEE, o desenvolvimento da inteligência artificial e sistemas autônomos tem levantado diversos problemas éticos, que podem ser compreendidos através do exemplo clássico do "The trolley problem", amplamente discutido por Thomson em 1985. Dilemas desse tipo têm demonstrado a necessidade de decidir o que é legalmente defensável quando um veículo autônomo se depara com uma situação de acidente que pode prejudicar seres humanos. Ou seja, certas decisões que seriam aceitáveis para um ser humano não necessariamente seriam toleradas pela sociedade quando tomadas por uma inteligência artificial ou incorporadas em sistemas de inteligência artificial. (IEEE, 2016)

Um exemplo recente que deve ser considerado é o problema enfrentado pela Amazon em 2015 com seu sistema de recrutamento baseado em inteligência artificial. A empresa percebeu que o sistema não estava classificando de forma neutra em termos de gênero os candidatos a empregos de desenvolvedor de software e outros cargos técnicos. Isso ocorreu porque os modelos de computador da Amazon foram treinados para examinar os currículos dos candidatos com base

em padrões de currículos enviados à empresa ao longo de uma década, em que a maioria dos candidatos eram homens devido ao domínio masculino na indústria de tecnologia. Consequentemente, o sistema de aprendizado de máquina acabou aprendendo que os candidatos do sexo masculino eram mais desejáveis, resultando em discriminação de gênero.(REIS, B. F.; GRAMINHO, V. M. C., 2019)

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.

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; NORVIG; DAVIS, 2010)

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, 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.

Diante do exposto acima, é demonstrado a relevância do tema de pesquisa, pois o desenvolvimento de sistemas inteligentes requer conhecimentos interdisciplinares, destacando-se a importância da área de ciências humanas na proposição de modelos que possibilitem o desenvolvimento de linguagens adequadas para a transmissão do pensamento e compreensão da moral por esses sistemas.

Com base na identificação desse problema epistemológico, estamos desenvolvendo uma tese que tem como objetivo argumentar sobre a possibilidade, ou não, de gerar conhecimento moral através da lógica para aplicação em sistemas inteligentes autônomos. Para tanto, pretende-se pesquisar as possibilidades de representação do conhecimento moral ou da moralidade através da lógica deôntica, lógica modal e da teoria dos modelos mentais de Johnson-Laird.

.

4. CONCLUSÕES

A moralidade é uma questão intelectual desafiadora, e com o avanço tecnológico, surge o debate sobre ensinar moralidade às máquinas. A hipótese geral é que a inteligência artificial estará cada vez mais envolvida em situações moralmente significativas. No entanto, não podemos atribuir responsabilidade moral aos sistemas autônomos apenas com base na lógica.

Partimos do pressuposto de que a sociedade ainda não possui padrões universais para incorporar normas e valores morais em sistemas inteligentes autônomos. Isso envolve formas altruístas e imparciais de cooperação, como em situações que exigem justiça.

Portanto, é essencial que os sistemas autônomos adotem as normas e valores morais da comunidade específica em que atuarão. Além disso, eles devem ser transparentes e confiáveis em suas ações.

O avanço tecnológico levanta questões éticas e morais, pois as tecnologias não são entidades autônomas na aquisição de dados, podendo esses dados conterem vieses humanos. É crucial que a comunidade acadêmica avance no estudo da ética no desenvolvimento de novas tecnologias e proponha modelos que permitam a transmissão do pensamento e da compreensão da moral aos sistemas autônomos.

5. REFERÊNCIAS BIBLIOGRÁFICAS

IEEE Institute of Electrical and Electronics Engineers. **A Iniciativa Global IEEE para Considerações Éticas em Inteligência Artificial e Sistemas Autônomos. Design com orientação ética: uma visão para priorizar o bem-estar com inteligência artificial e sistemas autônomos**, versão 1. IEEE, 2016. Online. Disponível em: http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

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

THOMSON, Judith. The trolley problem. **The Yale Law Journal**, 1985, v. 94, n. 6, p. 1395-1415

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