CC BY

VIRTUALIA JOURNAL
UFOP DEFIL Programa de Pós-graduação
hflps://periodicos.ufop.br/virtualia-journal
Editor responsável: Prof. Dr. Rodrigo Cid

Resenha de *Podemos Ensinar Moralidade às Máquinas? Inteligência Artificial e Ética em Debate* (2025)

Lucas de Azeredo Crespi

https://orcid.org/0000-0002-0122-1804
http://lattes.cnpq.br/3695477519778417
lucas.azecrespi@gmail.com

Resumo: nessa resenha é apresentada, analisada e discutida criticamente a obra *Podemos Ensinar Moralidade às Máquinas? Inteligência Artificial e Ética em Debate*, publicada em 2025 pela editora Dialética e escrita por Flávia Braga de Azambuja, doutora em filosofia pela UFPel. Derivada da tese de doutorado da autora, o escrito procura averiguar se uma Inteligência Artificial pode tomar decisões morais, o que implicaria a compreensão de conceitos como certo e errado, permitido e proibido, bem e mal. Para tanto, Azambuja investiga a possibilidade de representar, através da lógica modal, o abstrato conhecimento moral. O livro se destaca pelo teor interdisciplinar, pois lida com temas filosóficos, tecnológicos e inclusive psicológicos. O escrito é dividido em cinco segmentos, cada um com uma estrutura interna bem organizada e, em muitos casos, autossuficiente: a primeira parte introduz o campo da Inteligência Artificial e conceitos como *Machine Learning*, *Deep-learning*, Conexionismo e relacionados. Também na Parte um, Azambuja disserta sobre a filosofia moral consequencialista e utilitarista de John Stuart Mill, bem como a abordagem construtivista social de Jesse Prinz. Na segunda parte, a obra discute a lógica mental, os modelos mentais e o fenômeno do "viés da crença". O capítulo três é dedicado à lógica modal, um tipo de lógica formal não-monotônica que auxilia a capturar relações de possibilidade, necessidade e crença. Na quarta parte, as reflexões anteriores sobre lógica são utilizadas para lidar com três dilemas morais, que envolvem questões de saúde pública, liberdade individual e segurança nacional. A última parte do livro busca responder ao questionamento inicial, finalizando o argumento construído ao longo de toda a obra. *Podemos Ensinar Moralidade às Máquinas?* é louvável como obra introdutória e acessível. Todavia, por razões de cunho estrutural e metodológico, existem algumas insuficiências no tratamento dado ao assunto da moralidade, que perde espaço para a lógica e a psicologia cognitiva. A obra, portanto, se ocupa em demasia da filosofia teórica em detrimento da filosofia prática, o que se traduz numa resposta final que não é plenamente satisfatória diante do questionamento iniciado, considerando a amplitude de perspectivas e o intenso debate que é indissociável da filosofia moral.

Palavras-chave: Ética; Lógica; Conhecimento; Inteligência Artificial (IA); Psicologia cognitiva

Abstract: this review presents, analyzes and critically discusses the work *Can We Teach Morality to Machines? Artificial Intelligence and Ethics in Debate* (original title: *Podemos Ensinar Moralidade às Máquinas? Inteligência Artificial e Ética em Debate*), published in 2025 by Dialética and written by Flávia Braga de Azambuja, who holds a doctorate in philosophy from UFPel. Derived from the author's doctoral thesis, the text seeks to ascertain whether an Artificial Intelligence can make moral decisions, which would imply the understanding of concepts such as right and wrong, permissible and forbidden, good and evil. To this purpose, Azambuja investigates the possibility of representing abstract moral knowledge through modal logic. The book stands out for its interdisciplinary nature, as it deals with philosophical, technological, and even psychological themes. The text is divided into five segments, each with a well-organized internal structure and, in many cases, self-sufficient: the first part introduces the field of Artificial Intelligence and concepts such as Machine Learning, Deep-learning, Connectionism, and related terms. Also in part one, Azambuja discusses the consequentialist and utilitarian moral philosophy of John Stuart Mill, as well as the social constructivist approach of Jesse Prinz. In the second part, the work discusses mental logic, mental models, and the phenomenon of "belief bias." Chapter three is dedicated to modal logic, a type of non-monotonic formal logic that helps capture relations of possibility, necessity, and belief. In the fourth part, the preceding reflections on logic are used to address three moral dilemmas involving issues of public health, individual freedom, and national security. The last part of the book seeks to answer the initial question, finalizing the argument built throughout the entire work. "Can We Teach Morality to Machines?" is commendable as an introductory and accessible work. However, due to structural and methodological reasons, there are some insufficiencies in the treatment given to the subject of morality, which loses space to logic and cognitive psychology. The work, therefore, focuses too much on theoretical philosophy to the detriment of practical philosophy, which results in a final answer that is not fully satisfactory in view of the question posed, considering the breadth of perspectives and the intense debate that is inseparable from moral philosophy.

Keywords: Ethics; Logic; Knowledge; Artificial Intelligence (AI); Cognitive psychology

☐ **CRediT-IA (+):** Declaro que o uso de ferramentas de Inteligência Artificial foi estritamente instrumental e que assumo a responsabilidade humana integral pelo conteúdo do artigo. Simetricamente, autorizo o uso instrumental de ferramentas de Inteligência Artificial pelos pareceristas.
Ferramenta de IA utilizada:
Versão / modelo:
Finalidade do uso:
Tipo de uso:
Limites do uso:

☑ **CRediT-IA (-):** Declaro que não houve uso de IA para nenhum aspecto deste artigo e que assumo a responsabilidade integral pelo seu conteúdo. Simetricamente, não autorizo o uso de ferramentas de Inteligência Artificial pelos pareceristas.

Veja o modelo completo de Declaração CRediT-IA, criado pelo Virtualia Journal.

Introdução

Pode uma inteligência artificial pensar como os Homo *sapiens*? Relembremos o conto *Não tenho boca e preciso gritar* (ELLISON, 1967): um supercomputador chamado AM tem como desdobramento de sua tomada de consciência a gênese de um ódio mortal para com seus criadores, exterminando boa parte dos seres humanos no processo e se tornando uma espécie de déspota sádico e torturador, o que contradiz de maneira grotesca o seu propósito inicial, uma vez que AM se tratava de uma máquina destinada a proteger o território estadunidense de ameaças externas. O conto de Ellison, então, subverte o comportamento por hábito esperado de uma máquina, correspondente a submissão dócil e apática, colocando em seu lugar uma criatura com emoções negativas e sede de dominação. O próprio nome de AM é uma forma de referenciar o cogito cartesiano, visto que há um trocadilho entre AM – I am, eu sou – e a expressão latina *ergo sum* (ELLISON, 1967, p.3-4). A autoconsciência se torna palco para uma sátira cruel da raça humana, e não um tributo à sua racionalidade. Desse modo, não haveria correspondência necessária entre a consciência de si e a consciência moral. Pensemos nos animais não-humanos: muitos

são autoconscientes, mas isso tampouco implica em comportamentos morais. Golfinhos podem ser, de um ponto de vista humano, cruéis.

Pode uma inteligência artificial ser justa, então? Na série de animações japonesas *Psycho Pass* (2012-2023), escrita por Gen Urobuchi, não existe mais um sistema judiciário propriamente dito. Em seu lugar, há o chamado Sistema Sybil, uma IA sofisticada que, supostamente com base em um cruzamento de cálculos matemáticos, teorias jurídicas e filosóficas e dados de ponta, atua como juiz, júri e executor daqueles que identifica como criminosos. O Sistema Sybil, em conjunção com a força policial a ele subalterna, permite que armas especiais sejam disparadas para executar somente aqueles que são marcados com o rótulo de infratores em potencial. Apesar da eficiência manifesta desse método, é interessante refletir até que ponto ele é, como inquerimos, justo. É seguro delegar para um algoritmo a identificação de criminosos? Cabe a um sistema inteligente o dever que, até então, recairia sob o Estado e o pacto social que se presume sustentá-lo? Dados e informações brutas, por si só, não parecem suficientes para instituir justiça ou moralidade. Uma análise crítica de tais dados, portanto, é necessária. A pauta do racismo algorítmico[1], onde pessoas negras são equivocadamente tabeladas como criminosas, é uma exposição cruel dessa necessidade. O algoritmo pode identificar padrões e recorrência, mas não identifica nuances como a distinção entre correlação e causalidade, contexto histórico ou fatores econômicos e sociais. Dessa forma, a perpetuação de desigualdades e injustiças sistêmicas se esconde sob as vestes de um juízo

[1] Ver a obra *Algoritmos de Destruição em Massa* (O'NEIL, 2020).

técnico, científico e imparcial identificado com a máquina.

Pode uma inteligência artificial, finalmente, aprender a ser moral? Na coletânea de contos *Eu, Robô* (ASIMOV, 1950), conhecemos as Três Leis da Robótica e o seu impacto global no funcionamento desses seres:

> 1. Um robô não pode ferir um ser humano ou, por omissão, permitir que um ser humano sofra algum mal.
> 2. Um robô deve obedecer às ordens que lhe sejam dadas por seres humanos, exceto nos casos em que tais ordens contrariem a Primeira Lei.
> 3. Um robô deve proteger sua própria existência, desde que tal proteção não entre em conflito com a Primeira e a Segunda Leis.
> (ASIMOV, 1969, p.03).

Embora as Três Leis sejam de princípio claras e em plena harmonia, Asimov faz questão de jogá-las uma contra as outras no decorrer de alguns dos contos. A história "Runaround" é magnífica em demonstrar como a teoria é rapidamente desmontada pela prática, ao descrever um robô que vive num ciclo infinito de correr em volta de uma fonte de minerais radioativa para ele. Nesse caso, a Segunda Lei entra em conflito com a Terceira, e a máquina não pode obedecer plenamente a ordem de recuperar o mineral ao mesmo tempo em que protege a sua existência. Ser moral, de certo, requer um tipo de discernimento que vai além da simples e cega obediência a regras.

Seja em *Não tenho boca e preciso gritar*, *Psycho Pass* ou *Eu, Robô*, torna-se notável como a relação entre as máquinas e o senso de certo e errado consiste num tema de destaque na literatura de ficção científica. Com o avanço tecnológico, tais debates abandonaram o terreno meramente especulativo e adquiriram contornos mais e mais próximos do real. A crescente presença da inteligência artificial no cotidiano humano, seja em modelos de linguagem como o Chat GPT, assistentes virtuais como Cortana e Alexa e tantas outras variações exige diálogos que pautem os vínculos entre o virtual e o concreto,

questionando seus limites e possibilidades de aplicação prática. Nos dias de hoje, a atribuição de mais e mais ofícios inicialmente humanos para as máquinas provoca inquietude tanto nas massas quanto nos especialistas, de modo que a quantidade de perguntas supera em muito as escassas respostas disponíveis.

Nesse sentido, a filosofia se apresenta como uma ferramenta importante para desbravar algumas das tarefas em questão. A filosofia não constitui meramente o ato de pensar, mas uma atitude reflexiva onde o filósofo vai além disso e pensa sobre os próprios pensamentos. Tampouco constitui a simples busca do agir justo, mas dos fundamentos dos próprios conceitos de justiça, norma e ação. Acima de tudo, a filosofia se pauta na argumentação racional e, quando recorre a dados e teorias científicas, trata de pensar ambos em termos filosóficos.

É o que procura demonstrar a autora Flávia Braga de Azambuja, na tentativa de responder à pergunta que dá título à sua obra: *Podemos Ensinar Moralidade às Máquinas? Inteligência Artificial e Ética em Debate* (2025), publicada pela Editora Dialética. Azambuja é doutora em filosofia pela Universidade Federal de Pelotas (UFPel). A abordagem de Azambuja é educativa e excelente como tentativa de divulgação de conhecimento , mas a multiplicidade de assuntos abordados coexiste com algumas incompletudes em tópicos relevantes, tais como a filosofia moral e a filosofia política. O objetivo de Azambuja em sua obra é de caráter amplo. Conforme a autora:

> Este livro explora um dos dilemas mais intrigantes da atualidade: a capacidade (ou impossibilidade) da IA de tomar decisões morais. Investigamos se é possível programar uma máquina para agir de forma ética, se os algoritmos podem entender conceitos como empatia e justiça e quais os riscos de delegarmos decisões morais a sistemas automatizados (AZAMBUJA, 2025, p. 15).

É importante destacar que o livro é fruto de sua tese de doutorado, denominada *Sobre a viabilidade da representação de conhecimento moral em sistemas inteligentes: uma abordagem baseada em lógica não-monotônica* (AZAMBUJA, 2024). Assim, contrariamente à produção acadêmica que o gerou, o texto se propõe a ser compreensível para um maior número de pessoas, visto que sintetiza e simplifica ideias que são mais detalhadamente inquiridas no escrito de pós-graduação. Tanto o livro como a tese de doutorado, portanto, lidam com os mesmos temas, mas variam no grau de complexidade.

1. Estrutura da Obra

O livro é dividido em cinco partes ou capítulos. Na primeira delas, nomeada tal como o subtítulo do livro, ocorre uma introdução ao campo da inteligência artificial (IA) e seus principais conceitos: a autora expõe uma tabela de classificação de IA que leva em conta quatro critérios distintos: a) pensar humanamente; b) pensar racionalmente; c) agir humanamente; d) agir racionalmente, o que acentua uma certa dissociação entre ser humano e ser racional. Essa dissociação será importante no decorrer do livro. Em seguida, somos apresentados aos termos *Machine Learning* (aprendizado de máquina, ML) e *Deep-learning* (aprendizagem profunda, DL). Enquanto o ML se baseia em algoritmos para operar, sendo treinado com dados de modo a identificar padrões, correlações e pares input-output[2], o DL é intrinsecamente relacionado com uma abordagem denominada conexionista, na qual a IA

[2] Por exemplo: ao receber o input/entrada "o que é chocolate?", o algoritmo emite o output/saída "alimento baseado em cacau, de sabor ligeiramente amargo"

busca simular o cérebro e as suas redes neurais. O Chat GPT é uma manifestação de IA baseada em *Deep-learning*.

Além disso, na Parte um, Azambuja aborda as visões éticas de John Stuart Mill (2020) e Jesse Prinz (2007). É detalhado que Mill e Prinz defendem, respectivamente, teorias éticas baseadas no utilitarismo e no construtivismo cultural. Enquanto a abordagem de Mill considera que a moralidade é definida em termos de maior felicidade para o maior número, sendo as ações julgadas com base em suas consequências (AZAMBUJA, p.39); Prinz realiza uma crítica à ideia de que a moralidade seria uma faculdade inata ou um código universal, defendendo em seu lugar que a moralidade nasce da confluência entre alguns traços biológicos e experiências socioculturais do indivíduo, de modo que há nas normas morais um forte teor emocional e relativo.

Na segunda Parte, denominada "IA e Representação do Conhecimento Moral", a obra adentra a ideia de modelos mentais, entendidos como "representações internas simplificadas de sistemas externos para facilitar o raciocínio e a previsão de eventos futuros" (AZAMBUJA, 2025, p.57). Em outras palavras, na criação de um modelo mental, ocorre uma espécie de reconstrução do mundo e seus eventos dentro da mente, de maneira tal que o indivíduo possa enfrentar essa realidade na forma de preparo, análise passado-presente-futuro ou deliberação entre diferentes alternativas de ação possíveis. A lógica mental, evocada no início desse segundo capítulo, opera com base nos modelos mentais.

No capítulo seguinte, "IA e Representação Lógica do Conhecimento Moral", há novamente um mergulho na área da filosofia teórica, mais especificamente nos confins da lógica modal. A lógica modal é uma lógica formal não-monotônica, pois permite que "as conclusões não sejam

necessariamente preservadas quando novas informações são adicionadas" (AZAMBUJA, 2025, p. 76). Na lógica modal, há instrumentos para representar as noções de possibilidade (◊), necessidade (□) e crença (Bx), fundamentais para os conhecimentos abstratos como o conhecimento moral.

Na Parte quatro, "Dilemas Morais", as considerações prévias são utilizadas para tratar de três situações hipotéticas que serão representadas logicamente: a) pandemia de COVID e o Lockdown; b) legalização da *Cannabis sativa* e c) Armas de Destruição em Massa (ADM). Azambuja se vale, para tanto, da noção de experimentos mentais ou experimentos do pensamento[3]. Os dilemas morais presentes no livro são norteados por algumas regras básicas, que são simultaneamente responsáveis pela tensão interna inerente aos próprios experimentos mentais. No dilema sobre a pandemia e lockdown, há um atrito entre a necessidade de proteger a saúde e o bem estar da comunidade e, por outro lado, o risco de violações da autonomia e a intensificação de crises econômicas. Ao representar o dilema da legalização da maconha, são evidenciados pontos de conflito entre questões econômicas, justiça social e limites de ação do Estado. O dilema acerca das ADM é inspirado nos questionamentos do longa *Oppenheimer* (2023), de Christopher Nolan, e tensiona a busca por segurança nacional com a ameaça representada pelas bombas atômicas. Todos os três dilemas escancaram as capacidades e limitações da IA no que se refere a representar o conhecimento moral.

Finalmente, a última parte do livro responde à pergunta título, recapitulando as principais questões abordadas e realizando uma espécie de

[3] O argumento do gênio maligno, elaborado em *Meditações Metafísicas* (DESCARTES, 2016), é um experimento de pensamento. O mesmo ocorre com diversos argumentos contratualistas na filosofia política moderna, como a menção a um "estado de natureza" hipotético.

prognóstico crítico do tema "ética e Inteligência artificial" . A Parte 5 e suas conclusões serão analisadas posteriormente nessa resenha.

2. Aspectos positivos da Obra

Um dos maiores trunfos de "Podemos ensinar moralidade às máquinas?" é o seu método. A escrita segue uma trajetória escalonada e organizada por tópicos hierarquizados. Por exemplo: na Parte 2, começamos com uma introdução à teoria da lógica mental e suas limitações; em seguida, para aplicar a lógica mental, migramos para a teoria dos modelos mentais, e como ela pode ser utilizada para explicar o fenômeno do viés da crença, no qual "as pessoas tendem a aceitar conclusões que estão alinhadas com suas crenças, independentemente da validade real dessas conclusões" (AZAMBUJA, 2025, p.63). Logo, existe um senso de avanço conceitual, onde as lacunas do ponto A são preenchidas ou complementadas pelo ponto B. Os tópicos não são jogados ao acaso, mas interligados em uma dinâmica professoral que, assim como um algoritmo, segue uma sequência de passos.

Ademais, outro ponto de destaque na obra é a, até certo ponto, diversidade de conceitos e teorias utilizadas. Não há apenas o tratamento de áreas da filosofia e da inteligência artificial, mas também da psicologia cognitiva, especialmente no que se refere às ideias de Philip Johnson-Laird (2015). Esse intercâmbio só é possível pela temática em si mesma interdisciplinar, que converte o diálogo em um mandamento em vez de um reles conselho. A quarta parte do livro é exemplo desse caráter transversal, dado que conjuga lógica, ética e considerações de ordem política ou prática, como a regulamentação, as leis e o poder. Ao longo do livro inteiro, temos

contribuições tanto da filosofia analítica, com Searle (1980) e Quine (2011), como da filosofia moderna, com Hume (2009) e Leibniz (2023), havendo um cuidado notável em como cada uma dessas contribuições é empregada no texto.

Enquanto Searle, através do famoso exemplo do Quarto Chinês, auxilia a autora a dar, na Parte 1, considerações preliminares sobre a relação entre computadores e representação do conhecimento humano, Leibniz e sua teoria dos Mundos Possíveis[4] servem, na Parte 4, como uma excelente maneira de elucidar as categorias de possibilidade e necessidade presentes na lógica modal. No experimento do Quarto Chinês, Searle procura demonstrar que uma máquina não compreende de fato o idioma chinês, mas apenas manipula símbolos cegamente (sintaxe) sem atribuir significado (semântica) a eles. Esse gesto da máquina, enquanto carente de qualquer intencionalidade intrínseca, aponta para uma crítica à ideia de IA forte e, por extensão, ao teste de Turing[5]. O pensamento de Hume, análogamente às contribuições de Searle e Leibniz, ajuda a compreender o que pode ser definido como emotivismo moral, em contraposição a uma abordagem mais racionalista moral.

Azambuja realiza, com notável esmero, um escrito que vai além de uma colcha de retalhos ou mosaico de curiosidades. A caminhada investigativa e rigorosa se equilibra com os fins propedêuticos, sem que perca o seu espírito último. A escrita é o oposto da prolixidade ou hermetismo, sendo um exemplo notório de como existe a capacidade de tornar inteligíveis assuntos de início complexos. Não ocorre a presença de saltos lógicos ou

[4] Essa teoria é fundamental para a Teodiceia de Leibniz, satirizada em *Cândido* (VOLTAIRE, 2012).

[5] Logo, Searle procura demonstrar que é irrelevante se uma máquina nos convence de sua própria humanidade. Aparentar estar consciente de si, tal como a máquina aparenta ser falante de chinês, não implica necessária consciência ou, respectivamente, verdadeira compreensão do chinês.

grandes exigências de conhecimento prévio. Cada uma das cinco partes cumpre uma função bem delimitada para dar força ao argumento final da obra. Há diversos recursos que cumprem funções de esclarecimento e auxiliam o leitor, tais como tabelas, listas e subdivisões ou subtópicos. Deste modo, a absorção das informações presentes ao longo do livro ocorre tranquilamente.

3. Pontuando algumas dificuldades da Obra

No entanto, na mesma medida que a pluralidade de autores e teorias abordadas diversifica o conteúdo do livro, tal decisão criativa gera dois efeitos colaterais. A primeira dessas dificuldades é a redução do aprofundamento em certas áreas ou tópicos; o segundo é um efeito de dispersão que ocorre de modo concomitante com o ideal didático e metodológico previamente elogiado. Ou seja: ainda que a estrutura do livro seja excelente na forma e finalidade, nem todos os assuntos são aproveitados em seu máximo. Trata-se de uma limitação estrutural que não constitui um grande defeito per se, mas, a depender do leitor e suas predileções intelectuais, provoca uma sensação de carência e de falta.

Por exemplo, alguém que seja versado ou interessado em ética e filosofia moral verá, logo de início, que as considerações dessa ordem são inferiores em número de páginas quando comparadas com aquelas referentes a lógica, inteligência artificial e Psicologia cognitiva. Embora a ética permaneça como plano de fundo, considerando que o conhecimento que visa ter sua possibilidade de representação lógica investigada é justamente o abstrato conhecimento moral, ela é suprimida pela lógica e pela

epistemologia. Na obra de Azambuja, a filosofia prática é o eixo estruturante de uma cadeia composta por filosofia teórica e psicologia cognitiva. O fio condutor do livro é uma investigação que reside no âmbito da moral, mas os meios dos quais a autora se vale para avançar rumo a resposta para essa interrogação são provenientes de áreas mais distantes, ainda que potencialmente relacionáveis em algum grau. Nesse ínterim, cabe retornar à escolha dos autores-chave no segmento sobre filosofia moral. Pois, na Parte 5, com a finalidade de amparar a conclusão de que não é possível ensinar moralidade às máquinas, a autora argumenta em dois sentidos, se valendo dos pressupostos apresentados por Mill e Prinz: a) uma máquina pode se basear em dados e previsões para avaliar as consequências das ações, mas não pode compreender o que significa felicidade; b) as máquinas não tem emoções e c) as máquinas não podem ser motivadas moralmente. Nisso, conclui:

> 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 neste livro, 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 [...] (AZAMBUJA, 2025, p.121-122).

Ocorre que, a despeito da validade e plausibilidade do argumento, as preferências teóricas da autora levantam alguns questionamentos: sendo a filosofia moral um terreno tão diverso e repleto de debates, não seria adequado trazer mais autores e contrapontos? Pode-se trazer à mente a deontologia, a ética das virtudes - especialmente considerando os aportes presentes em *Depois da Virtude* (MACINTYRE, 2001) e *Modern Moral Philosophy* (ANSCOMBE, 1958) ou então o intuicionismo ético, expresso em

obras como *Ethical Intuitionism* (HUEMER, 2008) e *The Right and the Good* (ROSS, 2003). De modo geral, antes de questionar se podemos ensinar moralidade às máquinas ou se o conhecimento moral pode ser representado, Azambuja poderia ter refletido com mais profundidade sobre o que de fato é a moral.

Immanuel Kant seria, assim, um contraponto interessante aos ideais de Stuart Mill e Jesse Prinz. A ética kantiana, expressa em obras como a *Fundamentação da Metafísica dos Costumes* (KANT, 2009) e a *Crítica da Razão Prática* (KANT, 2016) é pautada por um ideal racionalista e defende que a moral não pode ser proveniente de fontes empíricas, tais como as emoções ou as normas socialmente construídas. Em segundo lugar, Kant é também um filósofo com uma veia muito crítica ao pensamento consequencialista. Para o intelectual prussiano, a moralidade não se encontra nas consequências ou nos impulsos sensíveis do agente, mas no motivo que mobiliza a sua vontade: uma boa vontade será determinada pelo senso de Dever racional, a despeito de quaisquer fatores empíricos. O último ponto, mais relevante, é o universalismo kantiano: as máximas das ações dos agentes, para que estes ajam moralmente, devem ser passíveis de universalização sem contradizer a vontade ou a lógica. Logo, por não admitir exceção, a filosofia moral kantiana seria excelente para refletir sobre a inteligência artificial. O elemento deontológico se assemelha àquele visto no princípio da resenha, na obra *Eu, Robô*. Embora Kant provavelmente não fosse considerar as máquinas como agentes morais[6], o recurso a sua filosofia iria conceder a Azambuja um panorama mais equilibrado do *status quaestionis*.

Na etapa conclusiva da obra é defendido que as máquinas não podem

[6] Para Kant, o agir moral pressupõe a liberdade da vontade e a sua não determinação pelo empírico, algo que as máquinas, por natureza, não possuem. IAs são determinadas, não livres.

ser encarregadas da tomada de decisões morais, necessitando de supervisão e intervenção humanas (AZAMBUJA, 2025, p.120). Embora essa postura derive logicamente das premissas da autora, existe um desconforto que as subjaz: seria esse interdito uma novidade ou consequência inesperada no campo? Isto é, não parece intuitivo que as máquinas não devam se envolver em áreas que tenham grandes implicações de ordem política ou ética? Não se trata de um apelo ao senso comum, mas do reconhecimento de que a relação entre a espécie humana e a tecnologia tende a consistir em um misto de entusiasmo pelo novo e a hesitação de ser dominado por ele.

Um dos maiores exemplos dessa dualidade entre o ser humano e a técnica ocorreu no passado: na Segunda Guerra Mundial, ou mesmo durante a maior parte do Século XX, os ideais do Iluminismo foram desafiados e a Europa se viu confrontada pelas consequências nefastas de sua própria racionalidade instrumental. Da esquerda à direita, da Escola de Frankfurt até a *Konservative Revolution*[7], uma ampla gama de pensadores destacou que a crítica da modernidade demanda uma crítica da Razão. Ou seja, ainda que hipoteticamente fosse possível internalizar nas máquinas um senso de moralidade, a hesitação em empoderá-las independe de qualquer apelo racional.

A Filosofia, como poderia ser dito numa paráfrase da décima primeira tese sobre Feuerbach (MARX e ENGELS, 2007, p.535), não deve se contentar em lidar passivamente com o mundo e o conhecimento: deve assumir uma postura ativa, crítica e transformadora em termos internos (a consciência, o conhecimento estabelecido como dominante) e externos (a sociedade, em

[7] Movimento ideologicamente heterogêneo e de influência sobretudo niedscheana, investigado em obras como *O modernismo reacionário* (HERF, 1993) e *The Conservative Revolution in the Weimar Republic* (WOODS, 1996)

suas dinâmicas de dominação). Nesse caso, ao discutirmos ética e moralidade, os dados e a lógica podem ser ferramentas úteis, mas não são critério suficiente quando isentos dessa atitude crítica. Azambuja, sem sombra de dúvidas, possui esse ímpeto questionador, mas infelizmente não o desenvolve em sua completude. Um capítulo extra seria muito bem vindo para a obra, de modo a balancear as considerações técnicas com as filosóficas e sociológicas. Sem esse cuidado, a conclusão da autora acaba interrompida no exato instante em que poderia se desenvolver rumo a algo mais complexo.

Considerações finais

Assim, *Podemos Ensinar Moralidade às Máquinas? Inteligência Artificial e Ética em Debate* (2025) é um livro que, acima de tudo, representa um esforço intelectual respeitável e é excelente enquanto modo de divulgação e introdução a temas importantes na Filosofia – especialmente a filosofia teórica -, psicologia cognitiva e estudos sobre inteligência artificial. Os atritos de ordem estrutural e pontualmente de conteúdo não suplantam, de maneira nenhuma, a capacidade de Azambuja de transpor para além da academia uma série de discussões que são tão difíceis quanto instigantes. É recomendado para acadêmicos ou entusiastas das ciências humanas, da programação ou de qualquer conjunto de temas que esteja aberto a uma investigação futurista. Pois, como demonstra a autora, a linha entre o ontem, o amanhã e o hoje é muito mais tênue do que gostaríamos de imaginar.

Referências

ANSCOMBE, Gertrude Elizabeth Margaret. Modern moral philosophy. *Philosophy*, v. 33, n. 124, p. 1-19, 1958.

ASIMOV, Isaac. *Eu, Robô*. Tradução de Luiz Horácio da Matta. 1969. Disponível em: https://kbook.com.br/wp-content/uploads/2016/07/eurobo.pdf. Acesso em: 12 de abril de 2026.

AZAMBUJA, Flávia Braga de. *Podemos Ensinar Moralidade às Máquinas? Inteligência Artificial e Ética em Debate*. São Paulo: Editora Dialética, 2025.

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.

ELLISON, Harlan. *Não tenho boca e preciso gritar*. 1967. Disponível em: https://notamanuscrita.com/wp-content/uploads/2015/06/visto-ellison-preciso-gritar.p df. Acesso em: 12 de abril de 2026.

ENGELS, Friedrich; MARX, Karl. *A ideologia alemã*. São Paulo: Boitempo, 2007.

HERF, Jeffrey. *O modernismo reacionário: tecnologia, cultura e política na República de Weimar e no Terceiro Reich*. Tradução de Cláudio Frederico Ramos. São Paulo: Ensaio, 1993.

HUEMER, Michael. *Ethical Intuitionism*. [s.l]: Springer Nature BV, 2008.

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

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

KANT, Immanuel. *Crítica da Razão Prática*. Petrópolis: Vozes, 2016.

KANT, Immanuel. *Fundamentação da Metafísica dos Costumes*. Tradução de Paulo Quintela. Lisboa: Edições 70, 2009.

LEIBNIZ, Gottfried Wilheim. *Ensaios de teodiceia*. [s.l]: WMF Martins Fontes, 2023.

MACINTYRE, Alasdair. *Depois da Virtude: Um Estudo em Teoria Moral*. 2. ed. [s.l.]: Editora da Universidade do Sagrado Coração, 2001.

MILL, John Stuart. *Lógica das ciências morais*. [s.l.]: Editora Iluminuras, 2020.

O'NEIL, Cathy. *Algoritmos de destruição em massa: Como o Big Data aumenta a desigualdade e ameaça a democracia*. Tradução de Rafael Abraham. Santo André: Editora Rua do Sabão, 2021.

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

QUINE, Willard Van Orman. *De um ponto de vista lógico - Nove ensaios lógico-filosóficos*. 1. ed. [s.l.]: Editora Unesp, 2011.

ROSS, William David. *The Right and the Good*. Oxford: Oxford University Press, 2002.

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

VOLTAIRE. *Cândido, ou o otimismo*. Tradução de Mário Laranjeira. [s.l]: Penguin-Companhia, 2012.

WOODS, Roger. *The Conservative Revolution in the Weimar Republic*. London: Palgrave Macmillan, 1996.