

Enviesamento algoritmo e improbidade intelectual no algoritmo Viola-Jones: o recurso genealógico como ferramenta para uma reflexão ética sobre os usos e aplicação da inteligência artificial (IA)

Felipe Cordeiro Alves¹

Resumo: Este trabalho utiliza premissas do procedimento genealógico nietzschiano para uma reflexão ética e crítica sobre a ausência de transparência no desenvolvimento da inteligência artificial e correlaciona essa situação com o estatuto da razão no interior da tecnociência. Nesse percurso investigativo é apresentada a tendência dos métodos usuais em *machine learning* em ocultarem os enviesamentos contidos no seu desenvolvimento, tendo como casuística o algoritmo Viola-Jones para reconhecimento facial. O método genealógico proposto por Nietzsche foi retomado como ferramenta para compreender como o desenvolvimento da inteligência artificial contempla, afirma e atinge diferentes grupos e a eticidade inaugurada por essa invenção. Com essa orientação, o estudo retoma as implicações das consequências indesejadas dos algoritmos de reconhecimento facial, sua dificuldade em reconhecerem faces de pessoas com a tonalidade de pele negra e como esse problema foi tardiamente reconhecido pelos desenvolvedores. Como saldo, a investigação apreende que o desenvolvimento da inteligência artificial ocorreu predominantemente orientado pela pressuposição da neutralidade da razão, bem como das funções intelectivas a ela relacionadas, contrariando as condições de seu exercício, que ocorre mobilizado por vontades e interesses parciais. Por fim, o trabalho indica o uso da inteligência artificial explicável como solução geral para o problema do enviesamento algoritmo.

Palavras-chave: inteligência artificial. Nietzsche. Genealogia. *machine learning*. Algoritmo.

INTRODUÇÃO

As aplicações práticas do aprendizado de máquina e as especulações sobre o potencial desse setor da tecnologia geram entre as coletividades uma espécie de fascínio misto, uma vez que são entendidos como fator decisivo para avanços tecnológicos e aprimoramentos inimagináveis para a população geral, ao mesmo tempo que podem atuar como fator de desestabilização da ordem social ao produzir processos como o desemprego estrutural e potencializar campanhas de desinformação.

Em face dessa incerteza, na qual a inteligência artificial promete cenários opostos, progridem mitologias sobre a inteligência artificial (IA), de forma positiva e negativa, que estão à disposição de mobilizações oportunistas. O eficaz retardamento dessa tendência passa por promover a acessibilidade e compreensão sobre esse ramo tecnológico, bem como absorvê-lo por meio de mediadores éticos, legais e culturais

A integração das IAs na consideração ética encontra entraves na pressuposição das máquinas como recursos desprovidos de viés, situação que exige o aprofundamento da reflexão sobre a eticidade de determinados processos do aprendizado das máquinas. Este trabalho

1 Mestre em Estudos Psicanalíticos (UFMG), Grupo Nietzsche (FAJE), felipepsi@live.com

recorre à ideia de probidade intelectual em Nietzsche e explora a pertinência das reflexões do autor para questões atuais, mostrando que o filósofo, contrariamente à reputação contraída nas discussões sobre a pós-modernidade, defende uma postura modesta em relação a nossas crenças e expectativas sobre o saldo da nossa atividade racional.

Por fim, o trabalho aponta preliminarmente para a convergência entre a proposta da inteligência artificial explicável e as considerações de Nietzsche sobre a probidade intelectual, compreendida como uma aceção da vontade de verdade [*Wille zur Wahrheit*], e o procedimento genealógico, encontrando nesse recurso um significativo instrumento para a absorção dos produtos da aprendizagem de máquinas por diretrizes comunitárias.

1 O ALGORITMO VIOLA-JONES

A ferramenta de detecção de objetos Viola-Jones foi desenvolvida por Paul Viola e Michael Jones com enfoque no problema de identificação de rostos, aplicação que foi posteriormente empregada para a detecção de objetos de modo geral. O recurso é baseado na identificação de características de Haar, a união de retângulos brancos e pretos em uma imagem para a identificação de texturas e outros padrões (MACEDO & GOMES, 2020). O algoritmo se divide em três partes: a criação de uma imagem integral, a montagem de um classificador de aprendizado *boosting* e a classificação em cascata (SANTANAS, GOMES e SANTOS, 2015). A primeira parte, a imagem integral, consiste em uma representação intermediária da imagem que contém a soma dos pixels delimitados por suas fronteiras (VIOLA & JONES, 2001). A segunda parte do algoritmo de reconhecimento responde pelo processo de *boosting*, no qual classificadores são aprimorados no contexto de escolhas binárias (MAYR, 2014). O recurso de *boosting* utilizado no algoritmo Viola-Jones é o *Adaboost*, que é direcionado para o estabelecimento de um classificador forte a partir de um classificador fraco ou de entrada (Schapire, 2003). Por fim, o terceiro passo, a classificação em cascata consiste na aplicação sequencial aos classificadores no qual as regiões que não apresentam pista de conterem o objeto de interesse são imediatamente descartadas e as regiões que contém os objetos procurados, no caso, rostos, são confirmados por múltiplos classificadores (MACEDO e GOMES, 2020).

Viola e Jones (2001), os idealizadores da ferramenta, deram acento à velocidade de processamento do recurso para detecção de rostos em seu trabalho inaugural. Essa característica, aliada à baixa exigência de processamento, são aspectos que permitiram o amplo emprego do algoritmo em diferentes situações e tarefas. A rápida resposta apresentada em aplicações, aliada à baixa exigência de recursos computacionais podem produzir a ilusão aos seus usuários de que a tecnologia é instantânea e unívoca na resolução de problemas.

Alguns anos após a proposta dos desenvolvedores, Schapire (2003) advertia sobre a dificuldade de classificadores finais desenvolvidas a partir do *Adaboost*, desenvolvido em 1995, serem compreendidos por seres humanos. Questionamentos recentes levantaram os vieses de raça e gênero nos empregos do algoritmo e decisões equivocadas obtidas por decisão algorítmica. Sandvig e colaboradores (2016) apontam que algoritmos apresentam um histórico

problemático no processamento de informação sobre aspectos raciais. Os autores advogam pelo estabelecimento de uma ética algorítmica, que tenha acesso às informações sobre o desenvolvimento e condições de aplicação dessas ferramentas.

Rudinskaya e Paringer (2020) avaliaram a acuidade das cascatas de classificadores Haar no reconhecimento de rostos, controlando as variáveis de gênero e raça. Os autores concluíram que o método de treinamento por cascatas Haar “podem ser posicionadas como detector universal, uma vez que não detectam bem o Homem negroide” (p. 240), incompletude que os estudiosos supõem ser decorrente da pobreza amostral de exemplares com essas características raciais nos treinamentos de máquina.

Em estudos mais recentes, Yucer e colaboradores (2023) revisaram o enviesamento racial no reconhecimento facial e encontraram indicativos de “preconceito racial em todas as etapas técnicas do processo” (p.29). Os investigadores também apontam para o aspecto geral dos acervos utilizados para treinamentos nos principais métodos de aprendizagem de máquina que não apresentarem diversidade racial, sendo que esse viés não foi, prosseguem o grupo de pesquisadores, objeto de um volume considerável de estudos sobre essa lacuna.

Retomando a especificidade do algoritmo Viola-Jones, temos como problemática as baixas exigências do recurso que obscurecem, na estimativa de custo-benefício, as dificuldades de sua aplicação e influência sobre acurácia de ambientes não controlados, distorção reforçada pela reputação positiva ostentada por inovações tecnológicas. Essa investigação desdobra a hipótese de que ocorre uma tendência, neste contexto ainda incipiente de aplicação da inteligência artificial, a dificuldade de reconhecer a parcialidade das decisões algorítmicas e reporta essa propensão ao estatuto atribuído à razão nas sociedades modernas.

O advento da inteligência artificial não se estabeleceu na observância de mediadores éticos e legais, de modo que sua aplicação ocorre de modo geral sem grandes considerações sobre esses dois domínios. Subjaz esse contexto de aplicação a crença na tecnologia e suas capacidades de predição, descrição e apreensão do mundo externo, situação sujeita a transformações que podem produzir a ampla aderência como a rejeição e demonização das ferramentas baseadas em inteligência artificial.

O caso dos métodos automáticos de detecção de notícias falsas pode servir de exemplar para as situações nas quais a crença nas potencialidades da máquina pode não comparecer. Os recursos de *learning machine* produziram algoritmos acurados na detecção de pseudonotícias a partir da notabilização da crise de desinformação, com diversos modelos chegando a uma eficácia superior a 90% (VILLELA *et al*, 2023). Contudo, esse desempenho não produziu uma reputação próxima da unanimidade entre as audiências, pois o ambiente de polarização em torno das questões que são impulsionadas pelas peças de desinformação, a parcialidade dos usuários e o contexto de questionamento da ciência estabeleceram a atuação de grupos significativos não aderentes à IA.

As aplicações obscuras do algoritmo Viola-Jones, sem esclarecer suas limitações e com excessiva confiança nas suas potencialidades para decisões capitais pode gerar grupos opostos à tecnologia e produzir modos de relação muito diferentes com a IA. Dessa forma, para que as ferramentas produzidas por esse campo não reafirmem preconceitos socialmente difundidos e que essas ocorrências não municiem manifestações anti-tecnologia e anticiência, faz-se necessária a observância de uma probidade intelectual cujas bases serão desenvolvidas nesta reflexão a partir do pensamento nietzschiano.

2 NIETZSCHE E A PROBIDADE INTELECTUAL

O recurso ao pensamento nietzschiano pode produzir estranhamento quando considerada a difundida associação do autor ao pensamento pós-moderno e a movimentos questionadores da ciência e *status quo*. Essa reputação dificulta a exploração das propostas de Nietzsche para o contexto contemporâneo e, por consequência, inviabiliza o reconhecimento da pertinência de seu pensamento para os desafios do nosso contexto.

É inegável a ocorrência de uma pronunciada tendência polemista, por vezes reconhecida pelo próprio filósofo, expressa em trabalhos como *Genealogia da Moral*, *O Anticristo* e em seu epistolário, do qual fazem parte as cartas da loucura. A notabilização desse aspecto do estilo de Nietzsche termina por obscurecer outros aspectos de seu projeto filosófico, entre eles a dimensão de probidade intelectual na vontade de verdade [*Wille zur Wahrheit*]. Para isolá-la, serão necessárias algumas mediações.

Podem ser depreendidos três principais sentidos para a noção de vontade de verdade na obra de Nietzsche, estando a primeira relacionada à vontade de obter a verdade referente a um mundo verdadeiro, estável e comandado por uma imutabilidade subjacente. Essa acepção da *Wille zur Wahrheit* pode ser apropriadamente definida como ilusão e corresponde à ilusão de um mundo verdadeiro metafisicamente fundado.

O segundo sentido é da vontade de verdade como tendência necessária para a existência e manutenção de modos de vida fracos, significado que dá a noção o estatuto de uma espécie de entorpecente. Os dois primeiros sentidos são reunidos como vontade de crença, de modo que a ilusão e a necessidade de ilusão ou de uma representação que ocupe de forma estável a representação da verdade figuram como uma forma da expressão de carência de forças. Esse par situa Nietzsche como um crítico da verdade e pensador contrário à ideia de verdade factual.

O terceiro sentido atribuído à vontade de verdade é o da probidade intelectual, escrípulo contido na pretensão de buscar a verdade superando os desvios instigados pela vontade de crença. De entrada, a probidade intelectual estabelece uma contradição performativa, visto que Nietzsche critica em sua obra a verdade ao mesmo tempo que o saldo da sua investigação sobre a verdade demite seu conceito orientador (JENKINS, 2012). A solução para esse aparente paradoxo é o entendimento de que a vontade de verdade corresponde ao impulso

ao veraz que apresenta por um lado a dimensão de correspondência com a realidade e, por outro, o de crença tratada como crença verdadeira.

Para Nietzsche não há um saber relevante para os interesses vitais que seja eterno ou facultade conhecedora e recurso que produza sempre verdades. Na ausência dessas garantias, o autor defende uma postura epistêmica modesta em relação às crenças e pretensas verdades. Assim, não há em Nietzsche uma oposição à verdade intelectual, o alvo do pensador são os dogmas e sistemas de pensamento dogmáticos cuja manutenção depende das tendências de ilusão e entorpecimento da *Wille zur Wahrheit*.

Outro aspecto assumido pela vontade de verdade como probidade intelectual é o da sinceridade. Esse sentido está relacionado à transparência sobre os motivos, impulsos e tensões que acompanham a produção do saber. Há especial interesse na obra de Nietzsche no desenvolvimento histórico das práticas morais e das crenças, inclinação do autor que levou ao desenvolvimento do procedimento genealógico.

O tipo de investigação histórica inaugurado pela genealogia teve como visada compreender, no contexto de investigação sobre a moral:

sob quais condições o seu humano inventou esses juízos de valor chamados de “bem e mal”? E qual o valor que possuem em si mesmos? Será que até agora reprimiram ou fomentaram o desenvolvimento humano? São um sinal de crise, de empobrecimento, de degeneração da vida? Ou, pelo contrário, expressam-se neles a plenitude, a força, a vontade para a vida, a coragem, a confiança nela, no futuro? (NIETZSCHE, 2017, p.27-28)

No marco teórico da Genealogia da moral, Nietzsche estava interessado em investigar a sucessão histórica das valorações “bem” e “mal”, abdicando de qualquer pressuposição histórico-finalista e voltado para o que considerava ser frutífero ou infrutífero para a cultura. Esse procedimento pode ser apropriado para a reflexão sobre a inteligência artificial e recursos associados à medida que a interrogação genealógica pode investigar os interesses, condições e preconceitos, descritos aqui como enviesamentos, que conduziram o desenvolvimento da inteligência artificial. Nessa reflexão preliminar, será apresentada a convergência das propostas para a inteligência artificial explicável com as concepções de probidade intelectual e do interesse genealógico em Nietzsche.

3 OPACIDADE, TRANSPARÊNCIA E PROBIDADE INTELECTUAL

Em face dos problemas e desafios produzidos por determinados usos de inteligência artificial, pesquisadores e trabalhadores especializados desenvolveram alternativas promissoras para seu enfrentamento. Conta entre essas propostas a inteligência artificial explicável que pode ser definida como processos de aprendizagem de máquina que seriam capazes de explicar suas bases, forças, fraquezas e suas tendências futuras.

Esse tipo de recurso tem como principal consequência superar a opacidade algorítmica, característica ainda predominante em algoritmos de *deep learning* que é tipificada por Nunes e Andrade (2023) em três setores, o sigilo intencional, analfabetismo técnico e complexidade algorítmica. A discussão proposta sobre o algoritmo Viola-Jones está centrada predominantemente no terceiro caso, podendo ser transposta, consideradas algumas mediações, para outros casos.

Reservadas considerações técnicas que não teriam lugar nessa reflexão, bem como considerações sobre arquiteturas de aprendizagem específicas que não seriam facilmente explicáveis por sua complexidade inerente, a proposta da inteligência artificial explicável apresenta convergência com as ideias de Nietzsche sobre a probidade intelectual e o projeto genealógico, uma vez que cumpre a virtude da sinceridade e fornece material para o procedimento genealógico.

Alves e Andrade (2021), sugerem que a explicabilidade seja um critério para aplicação da inteligência artificial em aplicações sensíveis, ou seja, em situações nas quais a explicabilidade ou opacidade seja relevante de entrada ou se prove subsequentemente um aspecto significativo. Cabe reforçar que a observância dessa transparência promete gerar impactos fundamentais sobre a veiculação dessas tecnologias, bem como a relação das instituições e populações seus setores produtivos.

A virtude da (im)probidade intelectual exercida pelos agentes dirigentes da produção e exploração de IAs são um aspecto que merece ser mencionado nessa reflexão, pois a opacidade dos algoritmos pode ocultar suas intenções ou serem resultado de omissão por parte dessas figuras. Certamente, boa parte das exigências sobre as IAs recairá sobre a classe empresarial detentora desses recursos. Do ponto de vista genealógico, o empresário das *Big Techs* representa uma versão do sacerdote ascético que alimenta de modo interessado a ideia tecnocrata de imparcialidade e de seus produtos como expressão do futuro. O futuro, pensando a partir de Nietzsche, dependerá do jogo de forças e de interesses no qual está envolvido a aplicação das IAs.

CONCLUSÃO

O algoritmo Viola-Jones foi tomado como exemplar de uma breve e preliminar exposição para apresentar as convergências entre as concepções éticas de Nietzsche sobre as práticas epistêmicas e seu procedimento genealógico. Esse percurso exigiu desassociar o filósofo da condição de pensador solidário às dificuldades geradas pela desinformação e discurso anticiência, demonstrando como o autor prima pela modéstia epistêmica e reativamente a concepções dogmáticas e juízos obtidos racionalmente. A transposição das reservas cautelosas do pensamento nietzschiano em relação à razão e ciência para o cenário de aplicação da inteligência artificial pode contribuir para a diluição da mitologização das IAs e reforçar o direcionamento das pesquisas e discussões públicas para debates construtivos, prevenindo que a implementação dessas tecnologias não seja influenciados pelas oscilações do imaginário

sobre essas tecnologias ou por grupos ou entidades oportunistas municiados pelos ruídos dos ambientes comunicativos.

REFERÊNCIAS

- ALVES, Marco Antônio Sousa; ANDRADE, O. M. de. Da “caixa-preta” à “caixa de vidro”: o uso da explainable artificial intelligence (XAI) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos, *Direito Público*, Brasília, v. 18, n. 100, 2022.
- JENKINS, Scott. Nietzsche's Questions Concerning the Will to Truth, *Journal of the History of Philosophy*, Charleston, v. 50, n. 2, p. 265-289, 2012.
- MACEDO, Anne L. F. & GOMES, Igor, R. Algoritmo Haar Cascade Aplicado na Detecção das Placas de Parada Obrigatória e de Velocidade Máxima Permitida, *XI Computer on the Beach, Camburiú*, v. 11, p.440-446, set/2020.
- MAYR, Andreas *et al.* *The evolution of boosting algorithms*, *Methods Inf. Med*, Stuttgart, v. 53, n. 06, p. 419-427, 2014.
- NIETZSCHE, Friedrich. *Genealogia da Moral* (Inês Lohbauer, trad), São Paulo: Martin Claret, 2017 (Trabalho original publicado em 1887).
- NUNES, Dierle José Coelho; DE ANDRADE, Otávio Morato. O uso da inteligência artificial explicável enquanto ferramenta para compreender decisões automatizadas: possível caminho para aumentar a legitimidade e confiabilidade dos modelos algorítmicos? *Revista Eletrônica do Curso de Direito da UFSM*, v. 18, n. 1, p. e69329-e69329, 2023.
- RUDINSKAYA, Elizaveta; PARINGER, Rustam. Face detection accuracy study based on race and gender factor using haar cascades, ITNT-2020, *Samara*, v. 1, n.1. p. 238-242, 2020.
- SANDVIG, Christian *et al.* When the Algorithm Itself Is a Racist: Diagnosing Ethical Harm in the Basic Components of Software, *International Journal of communication*, Los Angeles, v. 10, p. 4972-4990, 2016.
- SANTANAS, Luciana M; Q. SANTOS, Thiago, S. R; GOMES, Fábio R. Processo de detecção facial utilizando Viola-Jones, *Interfaces científicas*, Aracaju, v. 1, n. 1. p.35-40, fev/2015.
- SCHAPIRE, Robert, E. The Boosting Approach to Machine Learning: An Overview. In: *Nonlinear Estimation and Classification*. DENISON *et al* (Eds.), Berlim: Springer, 2003. p. 149-171.
- VILLELA, Humberto *et al.* *Fake news detection: a systematic literature review of machine learning algorithms and datasets*, Porto Alegre, v. 15, n. 1, p. 47-58, 2023.
- VIOLA, P. & JONES, M. *Rapid Object Detection using a Boosted Cascade of Simple Features*, *IEEE computer society conference on computer vision and pattern recognition*, Hilton Head, v. 1, p.511-518, 2001.
- YUCER, Seyma *et al.* Racial Bias within Face Recognition: A Survey, arXiv preprint arXiv:2305.00817, *Ithaca*, p. 1-35, mai/2023.