Principais Características para o Uso Responsável da IA

Daniela América da Silva[1], **Eduardo Wisnieski Basso**[2], **Johnny Cardoso Marques**[1]

[1] Instituto Tecnológico de Aeronáutica – São José dos Campos – SP – Brasil

[2]Universidade Federal do Rio Grande do Sul – Porto Alegre – RS – Brasil

damerica@ita.br, ewbasso@inf.ufrgs.br, johnny@ita.br

Abstract. *In recent years, the use of artificial intelligence (AI) has increased, also requiring intensive use of data. This scenario will require, in addition to current digital skills, the awareness to understand that data is never just raw, so we should not assume that the right types of AI will be developed and implemented on their own. This summary presents key traits to help in the best way to develop this technology plataform so present today.*

Resumo. *Nos últimos anos tem aumentado a utilização da inteligencia artificial (IA) requerendo também o uso intensivo de dados. Este cenário exigirá além das habilidades digitais atuais, a consciência para entender que os dados nunca são apenas brutos, portanto, não devemos assumir que, por conta própria, os tipos certos de IA serão desenvolvidos e implementados. Este resumo apresenta características chaves para auxiliar na melhor forma de desenvolver esta plataforma de tecnologia tão presente nos dias atuais.*

1. Introdução

Atualmente, a inteligência artificial (IA) vem sendo utilizada nas mais diversas áreas como sistemas de reconhecimento de voz, filtros de spam, sistemas de detecção de fraudes *online*, sistemas de recomendação de produtos, na educação, entre outras áreas e, as pessoas frequentemente assumem que os algoritmos são objetivos ou livres de erros. Porém quando usados em escala são propensos a serem implementados sem processo em vigor. Adicionalmente, as pessoas que trabalham com dados agem como intérpretes de fatos com significados ocultos e isso inclui seu próprio preconceito, trazendo um risco de viés de seleção.

Por exemplo, recentemente o MIT disponibilizou um Repositório de Riscos de IA [Slattery et al. 2024] com mais de 700 riscos potenciais que sistemas avançados de IA poderão representar. O estudo apresenta também que a maioria dos riscos da IA são identificados somente depois que um modelo se torna acessível ao público. E assim os criadores podem querer monitorar os modelos depois que eles são lançados, revisando regularmente os riscos que eles poderão apresentar após a implantação.

O objetivo geral deste trabalho é levantar questões sobre quais são as características necessárias, que poderão auxiliar na mitigação dos riscos em todo o ciclo de desenvolvimento da IA bem como durante a sua monitoração. Pois a IA é criada por humanos tendenciosos (como todos nós), e usar um modelo pode nos ajudar a minimizar o risco de preconceito em nossos projetos de IA. Primeiro, na seção Contexto, demonstramos a necessidade de comportamentos adequados e princípios para suportar

o desenvolvimento destes sistemas. Depois, a seção Modelo Proposto apresentará um modelo sobre como aplicar comportamentos, princípios e recomendações técnicas no desenvolvimento e monitoramento da IA. Finalmente, a seção Conclusão apresenta as principais realizações sobre o trabalho conduzido.

2. Contexto

Há uma série de coisas que vêm à mente sobre os diversos benefícios que a IA propicia, e ao mesmo tempo, há danos potenciais que surgem da IA, o que a torna uma área interessante para a ética [University of Melbourne 2023]. Recentemente o MIT disponibilizou um Repositório de Riscos de IA [Slattery et al. 2024], porém mesmo com o repositório é difícil saber com qual risco se preocupar mais, pois é necessário um melhor entendimento sobre como estes sistemas funcionam. Poderá haver também uma subjetividade envolvida na gestão de riscos, pois "os números são apenas ferramentas; eles não tem alma" [Bernestein 1996].

Um estudo do fórum econômico mundial [Ann Skeet 2020] revelou que 66% das pessoas se preocupam que a tecnologia tornará impossível saber se o que estão vendo ou ouvindo é real. Os consumidores também acreditam que a privacidade dos dados é crítica, e 53% dos consumidores globais pesquisados nunca usariam os produtos de uma empresa se seus dados fossem vendidos com fins lucrativos. Portanto, comportar-se eticamente definitivamente importa ao criar valor sustentável e as empresas podem criar valor duradouro para a sociedade alinhando suas práticas com as necessidades de todos os stakeholders, incluindo a comunidade em geral.

3. Modelo proposto

Embora haja um risco real ético, muitas vezes os problemas éticos em soluções de IA ocorrem pois pessoas bem-intencionadas falham em projetar a IA considerando de forma intencional o contexto ético. Desta forma é importante estabelecer e comunicar princípios simples que ajudem as pessoas a atingir o design ético. Para se adequar às características de seus usuários, as políticas que incentivam o comportamento ético devem, portanto, ser projetadas em torno de três processos psicológicos básicos que orientam o comportamento humano: atenção, interpretação e motivação. O estudo utiliza a ciência comportamental que identifica estas características como importantes para o design ético, e assim ajudam a ativar o comportamento ético [Epley and Tannenbaum 2017].

Adicionalmente, ver a tecnologia como "o herói" ou "o vilão" depende das escolhas que fazemos, pois as pessoas projetam e implementam a tecnologia e cabe a elas determinar quais limites serão colocados e definir o que conta como tecnologia "boa" e "ruim". Por isto os princípios éticos devem educar o design, desenvolvimento e implantação de novas tecnologias. E também podem servir como um padrão para testar se novas tecnologias passam no teste para encontrar defeitos e erros durante o processo de desenvolvimento bem como no monitoramento destas tecnologias [Matt Beard 2018].

Assim, este artigo propõe um modelo apresentado na figura 1 que considera que as políticas devem ser projetadas para ajudar as pessoas a manter os princípios éticos em mente, incentivar as pessoas a interpretar e entender as ramificações éticas de seu comportamento e, fornecer oportunidades e incentivos para perseguir objetivos éticos [Epley and Tannenbaum 2017] [Matt Beard 2018]. Duas apresentações foram feitas para

a equipe de dados como um treinamento interno de boas práticas, e para considerá-las ao construir o modelo.

3.1. Atenção - A ética está no topo do pensamento?

Pessoas éticas podem se comportar de forma antiética porque sua atenção está focada em outro lugar. Isto ocorre pois as pessoas têm atenção limitada e são guiadas por informações que são acessíveis, ou que estão na mente, no momento em que uma decisão é tomada. Um design de sistema eficaz leva as pessoas a pensarem sobre ética rotineiramente conforme os princípios de atenção adaptados de [Matt Beard 2018] e apresentados na tabela 1.

Tabela 1. Princípios de Atenção para Design Ético

Princípio	Descrição
Dever antes de poder. O fato de podermos fazer algo não significa que devamos.	Existem muitos mundos possíveis lá fora – muitas coisas que poderiam ser feitas ou construídas. O design ético é garantir que o que construímos ajuda a criar o melhor mundo possível. Antes de perguntar se é possível construir algo, precisamos perguntar por que utilizar IA em tudo.
Benefício Maximize o bem, minimize o mal.	As coisas que construímos devem dar uma contribuição positiva para o mundo - elas devem melhorar. Mas, mais do que isso, também devemos estar atentos aos efeitos colaterais potencialmente nocivos de nossa tecnologia. Mesmo que faça mais bem do que mal, o design ético exige que reduzamos ao máximo os efeitos negativos.
Responsabilidade Antecipar e projetar para todos os usos possíveis.	A tecnologia geralmente é projetada com um caso específico de uso e assim é possível prever as diferentes maneiras pelas quais as pessoas usarão nossos designs. Deixar de imaginar usos alternativos e suas implicações é arriscado pois poderá nos alertar sobre usos potencialmente nocivos contra os quais podemos nos proteger, ou benefícios que podemos maximizar.

3.2. Interpretação - Isto está correto?

O design eficaz ajuda as pessoas a reconhecerem a conduta ética e a ajustar o comportamento de acordo. Pois a maneira como as pessoas se comportam é influenciada pela maneira como elas interpretam — ou constroem — seu ambiente. Alterar a interpretação de um evento pode afetar dramaticamente o comportamento ao redefinir o que constitui conduta apropriada, conforme técnicas adaptadas de [University of Melbourne 2023] e apresentadas na tabela 2.

Tabela 2. Princípios de Interpretação para Design Ético

Princípio	Descrição
Coleta de dados. Os humanos decidem de quem coletar dados.	Quem é incluído e quem é deixado de fora? Quais etnias, em quais geografias ou jurisdições? Você só coleta dados de pessoas que parecem ter algum interesse no resultado final do sistema, ou de quem for conveniente?
Features. Os humanos decidem quais atributos têm mais importância e significado.	Que tipos de dados estão sendo coletados e quais não estão? Ignoramos ou coletamos gênero, idade, renda, estado civil, etnia ou código postal? Os dados de identificação devem ser coletados?
Projeto do algoritmo Os humanos fazem escolhas de design sobre algoritmos de IA.	Os falsos positivos devem ser minimizados ou os falsos negativos são mais importantes? Quantos dados são necessários para construir o modelo, etc.? Quais decisões são tomadas a cada passo, dando margem à possibilidade de introduzir viés?

3.3. Motivação - A solução está propiciando benefícios sociais?

As pessoas são motivadas por mais do que incentivos materiais - elas também têm motivações pró-sociais intrínsecas. O desejo de ajudar ou se conectar com os outros, pode ser usado para motivar comportamentos que se alinham naturalmente com práticas éticas. Em vez de focar em falhas éticas, as organizações devem chamar a atenção para comportamentos éticos exemplares para que outros imitem, conforme os princípios de motivação adaptados de [Matt Beard 2018] e apresentados na tabela 3.

Tabela 3. Princípios de Motivação para Design Ético

Princípio	Descrição
Não instrumentalismo. Nunca projete tecnologia na qual as pessoas sejam apenas parte da máquina.	Algumas coisas importam de maneiras que não podem ser medidas ou reduzidas ao seu valor de utilidade. Pessoas, ecossistemas, comunidades devem ser os beneficiários de seu design, não elementos de uma máquina ou design.
Propósito. Projete com honestidade e adequação de propósito.	O design requer ser honesto sobre a capacidade e limitações do projeto. Um projeto deve se destinar a resolver um problema genuíno. Um bom design serve a um propósito ético e o faz de maneira eficiente e eficaz.
Justiça. Trate casos semelhantes de maneira semelhante; casos diferentes de forma diferente.	Os projetos de tecnologia podem carregar vieses, e a justiça exige que apresentemos justificativas para quaisquer diferenças na forma como nosso design trata cada grupo de usuários. Devemos considerar se alguns grupos experimentam maiores danos ou menos benefícios do que outros.

Figura 1. Características Chaves para o Uso Responsável da IA

4. Conclusão

Este trabalho apresenta uma discussão sobre como conscientizar desenvolvedores de IA a construírem modelos com considerações éticas. E a partir da ciência comportamental é proposto um modelo com um conjunto de recomendações que seja relativamente simples de ser seguido no dia a dia da ciência de dados. Este estudo não se propõe a ser um fluxo de validação de ética, mas ajuda a mitigar riscos éticos no desenvolvimento de IA.

5. Agradecimentos

Os autores agradecem ao Grupo Boticário e ao Instituto Tecnológico da Aeronáutica pelo apoio geral para este trabalho.

Referências

Ann Skeet, Beena Ammanath, D. L. (2020). 5 traits of organizations that use tech responsibly. Technical report, World Economic Forum.

Bernestein, P. (1996). The new religion of risk management. MIT Technology Review.

Epley, N. and Tannenbaum, D. (2017). Treating ethics as a design problem. *Behavioral Science and Policy*, 3:72–84.

Matt Beard, S. L. (2018). Ethical By Design: Principles for Good Technology.

Slattery, P., Saeri, A., et al. (2024). The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence.

University of Melbourne (2023). Microcertification Introduction to the Ethics of Artificial Intelligence.