17º ERGODESIGN & USIHC 2019
PUC-Rio, 11 a 13 de dezembro
Rio de Janeiro, RJ, Brasil

17° Ergodesign – Congresso Internacional de Ergonomia e Usabilidade de Interfaces Humano Tecnológica: Produto, Informações Ambientes Construídos e Transporte
17° USIHC – Congresso Internacional de Ergonomia e Usabilidade de Interfaces Humano Computador

O Design pode ajudar na construção de Inteligência Artificial humanística?

Can Design help in the elaboration of a humanistic Artificial Intelligence?

CORTIZ, Diogo

PUC-SP e Ceweb.br , Doutor

diogocortiz@gmail.com

RESUMO

A inteligência Artificial (IA) atingiu um novo patamar de maturidade. Os sistemas já estão sendo empregados na área da saúde, educação e segurança. Mas será que eles levam em consideração valores humanos e culturais em suas decisões? Como o design pode ajudar a tornar a IA mais humanística? Neste artigo, apresenta-se os princípios para os projetos de IA e discutisse como o design pode ser aplicado em cada uma das etapas, apresentando um caso prático na elaboração de uma interface.

Inteligência artificial, design, princípios para inteligência, interface

ABSTRACT

Artificial intelligence (AI) has reached a new level of maturity. The systems are already in use in health, education and safety. But do they take human and cultural values into account in their decisions? How can design help make AI more humanistic? In this paper, we present the principles for AI projects and discuss how design can be applied in each of the stages, presenting a practical case in the creation of an interface.

Artificial intelligence, design, principles for AI, interface

1. A INTELIGÊNCIA ARTIFICIAL NA VIDA COTIDIANA

A Inteligência Artificial (IA) surgiu na década de 50 como uma área de estudo que buscava entender se a máquina seria capaz de imitar a cognição humana. Podem as máquinas pensar? É com esta pergunta em um artigo científico que Alan Turing (TURING, 1950) inaugura um ramo da Ciência da Computação que se ocupa com pesquisas sobre técnicas, mecanismos e dispositivos que possam aplicar maneiras similares como aprendemos e tomamos decisões. Apesar de muitos pesquisadores considerarem este artigo como um dos pontos de partida para os estudos técnicos, filosóficos e sociais da IA, é importante destacar que Turing nunca utilizou este termo para descrever as suas inquietações sobre a área.

O termo Inteligência Artificial, que hoje representa uma grande área, foi utilizado pela primeira vez em 1956, durante *The Dartmouth Summer Research Project on Artificial*

Intelligence, um evento seminal que reuniu um grupo seleto de pesquisadores renomados, como John McCarthy, Marvin Minsky, Claude Shannon, que, além de ser um espaço inicial para se pensar o que seria a inteligência e quais os princípios que uma máquina poderia seguir para imitá-la, também acabou criando o termo que se fortaleceu ao longo das décadas (BUCHANAN, 2005).

A história da IA foi marcada por momentos de expectativas e de esquecimentos, com idas e vindas do interesse da comunidade acadêmica e financiamento em pesquisas. Para se ter uma ideia desta dinâmica, podemos analisar o cenário do desenvolvimento tecnológico entre os anos 90 e 2000, um momento de muitas especulação e promessas, em que as oportunidades e investimentos estavam voltadas especialmente para as empresas que lançavam serviços para a Internet. Muitas empresas que hoje lideram pesquisas e serviços na área de IA, como Google, Facebook e Amazon, não foram criadas com o propósito de desenvolver tecnologias para esta área, mas hoje possuem uma posição privilegiada e oferecem os mais diferenciados tipos de serviço de IA – tradução, comando por voz e reconhecimento facial, por exemplo – porque conseguiram se beneficiaram de três principais fatores: a abundância de dados, a evolução do poder de processamento e o surgimento de novas abordagens e técnicas.

Na última década, houve um aumento na acurácia e no desempenho dos sistemas inteligentes, e, aos poucos, a Inteligência Artificial passou a mostrar resultados em escala global. Hoje os algoritmos de IA estão presentes em diversos serviços disponíveis, de sistemas de recomendação de música aos sistemas antifraude do cartão de crédito. No entanto, ainda existem muitas dúvidas sobre o que é realmente a Inteligência Artificial e o que ela será capaz de fazer e resolver no futuro. No escopo deste artigo, nos atentaremos ao estado da arte da IA, suas aplicações e limitações. Neste sentido, o nosso entendimento para fins desta pesquisa é o de que a Inteligência Artificial se limita às técnicas de Aprendizado de Máquina para solucionar problemas específicos.

As técnicas comumente utilizadas no Aprendizado de Máquinas se dividem em duas principais abordagens: o aprendizado supervisionado e não-supervisionado. Na primeira, os sistemas observam exemplos de pares de entrada e saída e aprendem uma função capaz de mapear uma nova entrada para uma saída (RUSSELL; NORVIG, 2009). Isso acontece a partir de um grande volume de dados rotulados, ou seja, dados que tenham sido previamente “explicados”. Por exemplo, se o objetivo for treinar um algoritmo de classificação de imagens de frutas, serão necessárias centenas de milhares imagens de frutas para treiná-lo, e cada imagem devem conter as informações, ou seja, “os rótulos”, sobre o que representam: maçã, banana, uva.

Assim, por meio de diversos exemplos, o sistema aprende que existem diferentes formatos, cores, texturas para maças. O aprendizado supervisionado permite que desenvolvamos sistemas de classificação e regressão, como reconhecimento facial, predição de preços, entre outros.

Já as técnicas de aprendizado não-supervisionado não demandam que os dados estejam obrigatoriamente rotulados, ou seja, “explicados”. Elas são capazes de criar agrupamento e segmentações que nos permite entender melhor os próprios dados. Imagine um banco que tenha uma grande quantidade de dados de transações de compras que nunca tenham sido rotuladas como fraude ou não-fraude. Neste caso, não será possível que um sistema aprenda o que é ou não é fraude, mas por meio de técnicas de aprendizado não-supervisionado o sistema poderá aprender os padrões das transações para agrupá-las pelo mesmo perfil. Essa abordagem já é utilizada por empresas para segmentar seus clientes com o mesmo padrão de compra e usuários de um serviço com o mesmo gosto e preferência, por exemplo.

Um algoritmo bem treinado é capaz de reconhecer padrões que são invisíveis para nós. As técnicas de aprendizado de máquina identificam padrões que muitas vezes são difíceis até mesmo para um especialista. Por exemplo, uma pesquisa mostrou que técnicas de *deep learning* podem prever Alzheimer 6 anos antes do diagnóstico médico. Isso é possível porque este tipo de algoritmo identifica padrões de mudanças sutis e globais no cérebro, a partir de imagens de PET Scan, o que é difícil para um radiologista humano (DING *et al.*, 2019).

Como visto, as principais técnicas de aprendizado de máquina dependem de dados para que aprendam a partir de exemplos. No entanto, é preciso ter cuidado com a fonte e qualidade dos dados desde o início, para que a os algoritmos sejam treinados da melhor maneira possível. A forma como escolhemos os dados influenciará em como o sistema aprenderá e se comportará - o que mais uma vez reforça a necessidade de trabalhar corretamente esses dados, já que as decisões tomadas pelos algoritmos e máquinas podem gerar impactos em larga escala.

Um estudo publicado na revista Science (Obermeyer et al., 2019) demonstrou que um sistema usado para gerenciar a saúde de pacientes em um hospital nos Estados Unidos apresentava um possível comportamento de discriminação racial. O software em questão era o responsável por alocar atendimentos personalizados e tratamentos especiais para pacientes com doenças crônicas, mas a análise de suas decisões mostrou uma tendência de privilégios aos pacientes brancos quando comparado aos negros. Dado dois pacientes classificados da com a mesma prioridade para receber o tratamento especial, os pesquisadores notaram que os negros geralmente estavam mais doentes.

Essa situação aconteceu porque a aplicação foi desenhada para que o contexto da decisão fosse baseado na tendência de gastos para o futuro no tratamento do paciente. No entanto, sabe-se que os algoritmos são treinados com dados do passado, e que os negros têm históricos de limitações de acesso ao sistema de saúde com gastos menores do que os brancos em média. Em outras palavras, o sistema aprendeu que os pacientes brancos teriam um gasto maior, portanto seria melhor alocá-los em um atendimento preferencial. Mas por se tratar de um sistema de alocação de tratamento para pacientes com base no nível de sua doença, as variáveis decisórias usadas no projeto do sistema talvez não devessem ser financeiras, mas focadas apenas no estado de saúde de cada paciente.

Situações similares acontecem em diversos contextos nos quais algoritmos de aprendizado de máquina são implementados. Um caso seminal objeto de estudo na comunidade acadêmica foi a investigação conduzida pela agência de jornalismo ProPublica, que mostrou que o sistema utilizado para o cálculo de reincidência criminal também apresentava resultados tendenciosos para os negros (ANGWIN *et al.*, 2016). A discriminação não se limita apenas a critérios raciais, há casos de sistemas inteligentes com comportamentos discriminatórios também por causa de religião, sexo, idade ou renda.

Esses são alguns dos desafios para as próximas décadas para que possamos construir sistemas de Inteligência Artificial que levem em consideração os direitos e valores humanos. Debates estão acontecendo em governos, organizações internacionais e empresas de tecnologia. Neste artigo serão apresentados os principais princípios para uma IA humanística para que possamos discutir como o design – e suas abordagens, métodos e técnicas - possam auxiliar neste processo.

2. PRINCÍPOS PARA INTELIGÊNCIA ARTIFICIAL

O desenvolvimento acelerado da Inteligência Artificial nos últimos anos está chamando a atenção não apenas da comunidade técnica e acadêmica, mas também de governos, instituições sociais e organizações internacionais que estão preocupadas com os desdobramentos das tecnologias na sociedade. Conforme visto anteriormente, um algoritmo treinado de forma inadequada e sem as devidas validações pode causar danos e prejuízos para um grupo de pessoas, uma etnia ou uma raça.

Recentemente, o tema ganhou espaço em debates que acontecem em diferentes fóruns – dos mais técnico aos mais políticos. E muitas instituições começaram a produzir e divulgar materiais com princípios básicos para projetos de IA. Para este artigo, foi feito um levantamento dos principais documentos para que possamos ter um panorama dos principais princípios propostos para um debate global sobre IA. Essa é uma parte importante do

processo para que possamos entender o que se espera de uma IA humanística para que, então, possamos discutir sobre como o design pode auxiliar nesta transformação. Para o escopo deste artigo, utilizamos os seguintes documentos:

- *AI Principles: Recomendations on the Ethical Use of Artificial Intelligence by the Department of Defense* – documento publicado pelo Departamento de Defesa que traz os princípios e recomendações éticas para o uso de IA na área da defesa (UNITED STATES DEPARTMENT OF DEFENSE, 2019).

- *High-level Expert Group on Artificial Intelligence – Ethics guidelines for trustworthy AI* – documento organizado pela Comissão Europeia que elabora um *framework* para uma IA que seja confiável, além de propor princípios para o desenvolvimento e implementação da tecnologia (EUROPEAN COMMISSION, 2019).

- *Recommendation of the Council on Artificial Intelligence* – documento produzido pela Organização para a Cooperação e Desenvolvimento Econômico para que possa guiar o processo do desenvolvimento de uma inteligência artificial humanística nos países integrantes(OECD, 2019).

- *AI at Google: our principles* – página disponibilizada pelo Google em que o presidente da organização comunica os princípios que deverão guiar o desenvolvimento de projetos de IA dentro da multinacional (PICHAI, 2018)

- *Microsoft AI Principles* – página disponibilizada pela Microsoft que publica os princípios de IA na organização (MICROSOFT, 2019).

- *Beijing AI Principles*– documento produzido por um consórcio chinês de universidades e principais empresas de tecnologias para guiar o desenvolvimento ético da IA no país(“Beijing AI Principles”, 2019).

Selecionamos estes documentos na busca de garantir uma pluralidade não só regional (Estados Unidos, Europa e Ásia) – não foi encontrado nenhum material produzido para o Brasil até o momento da escrita deste artigo – como também por área de atuação (setor privado, governos e organizações internacionais). Após a leitura de cada documento, mapeamos os princípios comuns em cada documento, o que nos permitiu criar dimensões de princípios para que pudéssemos entendê-los melhor. As dimensões que chegamos após esta etapa estão resumidas abaixo:

- Dimensão de *Fairness*: dimensão em que agrupamos os princípios que abordam questões como não discriminação, a não violação de direitos humanos e a inclusão de valores democráticos, de inclusão e a eliminação de vieses indesejados.

- Dimensão de Confiabilidade e Segurança: dimensão em que agrupamos os princípios que defendem que o sistema tenha sido construído e testado para garantir a segurança física e social de seus usuários, assim como garantir que salvaguardas sejam criadas para evitar grandes impactos quando uma situação indesejada ocorrer.

- Dimensão de Responsabilidade: dimensão em que agrupamos os princípios que citam que os sistemas de IA devam ter responsabilidade algorítmica e que os desenvolvedores considerem de potenciais riscos éticos, sociais e econômicos.

- Dimensão de Privacidade: dimensão em que agrupamos os princípios relativos à proteção da privacidade dos usuários e a garantia de liberdade e autonomia.

- Dimensão de Transparência: dimensão em que agrupamos os princípios que buscam fortalecer a necessidade de que os sistemas sejam transparentes e suas decisões explicadas, ainda que esse ainda seja um desafio técnico contemporâneo

O mapeamento dos princípios e a elaboração destas dimensões são importantes porque nos deixa um panorama dos alicerces para uma Inteligência Artificial humanística. Entendemos que apesar de uma amostra limitada no número de atores mapeados, o processo contemplou importantes instituições que estão liderando e influenciando este debate na cena internacional.

É importante ressaltar também que os princípios devem ser entendidos como um conjunto de guias e normas a serem seguidos, o que pode ser um complicador na área da Inteligência Artificial devido a sua complexidade técnica. Muitos princípios são importantes, mas ainda não se sabe muito bem como atendê-los na implementação dos modelos. Durante décadas, os estudos de inteligência artificial ficaram concentrados nas áreas de ciência da computação, matemática, engenharia e ciências cognitivas, mas com o seu estágio atual de maturidade e de aplicações na sociedade, percebe-se a necessidade de envolvimento de um grupo mais diverso e interdisciplinar para trazer novas perspectivas e conhecimentos aos projetos. E o Design, com suas abordagens, técnicas e ferramentas, pode ajudar para que esses princípios sejam considerados em projetos de Inteligência Artificial em diversas etapas do projeto.

3. AFINAL, O DESIGN PODE AUXILIAR NOS PROJETOS DE UMA IA HUMANÍSTICA?

Neste artigo, argumentamos que o Design é uma área que deve estar envolvida em todas as fases de um projeto de IA para que valores humanos culturais possam ser mapeados, entendidos e incorporados ao sistema. Conforme visto anteriormente, os algoritmos de IA aprendem a partir de dados, mas muitas vezes esses conjuntos de dados podem conter alguns vieses da sociedade. Entendemos que há técnicas de design que possam ser aplicadas para a criação de métricas para entendimento dos dados bem como propor uma interface mais inclusiva ao usuário.

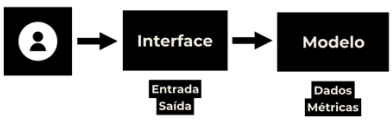

Figura 1 – Esquema de Interação com um sistema de IA

A Figura 1 demonstra o esquema de interação do usuário com um sistema de inteligência artificial. Observa-se que o usuário final não tem acesso ao modelo treinado, muito menos aos dados que foram utilizados no treinamento. A interação se dá principalmente por meio da interface. Apesar de argumentarmos que o design pode ser ajudado em todas as etapas do projeto de IA, neste artigo, o recorte é sobre como o design pode ajudar na construção das interfaces para garantir os princípios de *Fairness* e Confiança. Apesar de entendermos que existam outras técnicas de design que ajudam na construção de métricas para a construção de modelo, esta discussão não será contemplada neste artigo.

Trazemos um exemplo prático identificado no ano de 2019. Após uma investigação sobre representatividade e gênero em sistemas de processamento de linguagem natural e tradução, foi identificado que a aplicação de tradução do Google (*Google Translator)* apresentava um viés de gênero ao traduzir do inglês para o português as palavras "*doctor" e "nurse".* O sistema traduzia de uma forma automática a palavra "*doctor"* como "médico" e "nurse" como "enfermeira". Não havia nenhuma opção para traduzir "*doctor*" como "médica" ou "*nurse*" como "enfermeiro". Conforme pode ser observado na Figura 2, o usuário não tinha opção de escolha.

Figura 2 – Captura do sistema de tradução do Google em janeiro de 2019

Neste caso, o problema não estava apenas nas interfaces do sistema de tradução, porque apenas exibiam o que o modelo determinava. Também é preciso notar que o próprio modelo não foi desenvolvido propositalmente para se comportar assim, mas ele aprendeu a partir de exemplos – textos disponíveis na internet e livros, por exemplo – que existiam mais mulheres enfermeiras do que homens enfermeiros, assim como existiam mais homens médicos do que mulheres médicas. Essa é uma explicação simples e resumida de como o algoritmo na área de processamento de linguagem natural aprende para que não tenhamos que entrar em detalhes técnicos com o risco de fugir do escopo do trabalho. Mas ao adotar essa regra, o sistema ignorava a possibilidade de existência de uma mulher médica e de um homem enfermeiro.

Esta situação parece desrespeitar pelo menos duas dimensões dos princípios mapeados: a dimensão de *fairness* e a dimensão de confiabilidade. A primeira está relacionada com o fato de que o sistema apresentou um comportamento discriminatório, não tratando de maneira igualitária os gêneros nas possibilidades de suas profissões – assumindo que uma mulher só poderia ser enfermeira enquanto o homem só poderia ser médico. A segunda está relacionada com a falta

de confiança que esta situação causava aos usuários, uma vez que eles não tinham controle para fazer a escolha do genero que achassem mais apropriado.

A partir de maio de 2019, identificamos que uma nova interface ajudou a solucionar este problema. O sistema passou exibir ambas opções para o usuário. Não é possível saber se também houve alteração nos modelos de aprendizado de máquina, mas a nova versão da interface, conforme pode ser observado na Figura 3, é mais adequada para um sistema de tradução que deseja evitar perpetuar vieses e discriminações.

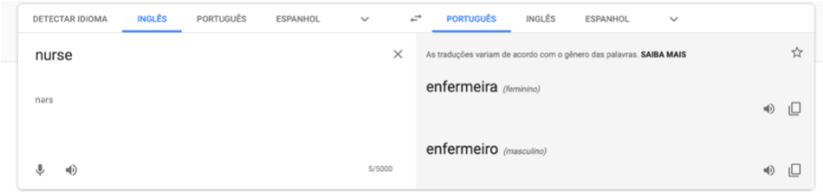

Figura 3 – Captura da nova versão do sistema de tradução do Google

Na nova versão, o sistema de certa forma consegue sustentar os princípios de *Fairness,* ao não discriminar uma profissão por gênero, como acontecia na versão anterior, e também os princípios da dimensão de Confiabilidade, porque agora o usuário tem mais controle sobre o sistema a partir da liberdade de escolha. Este caso demonstra que o design pode auxiliar na construção de interfaces que ajudem a suportar os princípios para uma inteligência artificial humanística, mas reforçamos que está é uma área em que durgem técnicas e abordagens para ajudar também no entendimento do modelo e nas métricas dos dados, assuntos que serão tratados em trabalhos futuros.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ANGWIN, J. *et al.* Machine Bias. *ProPublica*, 2016.

Beijing AI Principles. Disponível em: <https://www.baai.ac.cn/blog/beijing-ai-principles>. Acesso em: 25 nov. 2019.

BUCHANAN, B. G. A (Very) Brief History of Artificial Intelligence. *AI Magazine*, v. 26, n. 4, p. 53–53, 15 dez. 2005.

DING, Y. *et al.* A deep learning model to predict a diagnosis of Alzheimer disease by using 18 F-FDG PET of the brain. *Radiology*, v. 290, n. 3, p. 456–464, 1 mar. 2019.

EUROPEAN COMMISSION. High-Level Expert Group on Artificial Intelligence. p. 2–36, 2019. Disponível em: <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top>.

MICROSOFT. *Microsoft AI Principles*. Disponível em: <https://www.microsoft.com/en-us/ai/our-approach-to-ai>. Acesso em: 21 nov. 2019.

OBERMEYER, Z. *et al.* Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, v. 366, n. 6464, p. 447–453, 25 out. 2019.

OECD. *Recommendation of the Council on Artificial Intelligence*. Disponível em: <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>.

PICHAI, S. *AI at Google: our principles*. Disponível em: <https://www.blog.google/technology/ai/ai-principles/>. Acesso em: 22 nov. 2019.

RUSSELL, S. J.; NORVIG, P. *Artificial intelligence a modern approach*. 3rd. ed. New Jersey: Prentice Hall, 2009.

TURING, A. M. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, v. LIX, n. 236, p. 433–460, 1 out. 1950. Disponível em: <https://academic.oup.com/mind/article/LIX/236/433/986238>. Acesso em: 21 nov. 2019.

UNITED STATES DEPARTMENT OF DEFENSE. AI Principles : Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense Defense Innovation Board Supporting Document. p. 1–66, 2019.