UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE QUIXADÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

MESTRADO ACADÊMICO EM COMPUTAÇÃO

NILTEMBERG DE OLIVEIRA CARVALHO

MOREXAI: UM MODELO PARA REFLETIR SOBRE INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

QUIXADÁ

2022

NILTEMBERG DE OLIVEIRA CARVALHO

MOREXAI: UM MODELO PARA REFLETIR SOBRE INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

Dissertação apresentada ao Curso de Mestrado Acadêmico em Computação do Programa de Pós-Graduação em Computação do Campus de Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Engenharia de Software.

Orientadora: Profa. Dra. Andréia Libório Sampaio.

QUIXADÁ

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

C326m Carvalho, Niltemberg de Oliveira.
MOREXAI: Um modelo para refletir sobre Inteligência Artificial Explicável / Niltemberg de Oliveira Carvalho. – 2022.
94 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Campus de Quixadá, Programa de Pós-Graduação em Computação, Quixadá, 2022.
Orientação: Profa. Dra. Andréia Libório Sampaio.

1. Engenharia Semiótica. 2. Inteligência Artificial. 3. Ética. I. Título.

CDD 005

NILTEMBERG DE OLIVEIRA CARVALHO

MOREXAI: UM MODELO PARA REFLETIR SOBRE INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

Dissertação apresentada ao Curso de Mestrado Acadêmico em Computação do Programa de Pós-Graduação em Computação do Campus de Quixadá da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Engenharia de Software.

Aprovada em: __/__/__.

BANCA EXAMINADORA

Profa. Dra. Andréia Libório Sampaio (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Davi Romero de Vasconcelos
Universidade Federal do Ceará (UFC)

Prof. Dra. Ingrid Teixeira Monteiro
Universidade Federal do Ceará (UFC)

Profa. Dra. Marcelle Pereira Mota
Universidade Federal do Pará (UFPA)

À minha família, por sua capacidade de acreditar em mim e investir em mim. Mãe, seu cuidado e dedicação me deram, em alguns momentos, a esperança para seguir.

AGRADECIMENTOS

À minha orientadora Profa. Dra. Andréia Libório Sampaio, por sua dedicação, apoio e por sempre ter acreditado e depositado sua confiança em mim ao longo dessa trajetória, e que já me acompanha desde a graduação.

Aos professores participantes da banca examinadora Prof. Dr. Davi Romero de Vasconcelos, Prof. Dra. Ingrid Teixeira Monteiro e Profa. Dra. Marcelle Pereira Mota pelo tempo, pelas valiosas colaborações e sugestões desde a qualificação.

À minha querida mãe, a quem dedico esta dissertação, pelo apoio incondicional e pelos valores que sempre me transmitiu, dentre os quais a força para nunca desistir de lutar. Ao meu marido, Vinicius Scheffer, cujo apoio foi essencial. Aos meus amigos que sempre estão comemorando minhas vitórias e me dando força nos momentos difíceis.

Aos órgãos de fomento à pesquisa. À Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico, pelo apoio financeiro com a manutenção da bolsa do Governo Digital. O trabalho foi realizado também com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

"Se você não pode explicar algo de forma simples, então você não entendeu muito bem o que tem a dizer!" (Albert Einstein)

RESUMO

O interesse por sistemas que utilizam aprendizado de máquina vem crescendo nos últimos anos. Alguns algoritmos implementados nesses sistemas inteligentes ocultam seus pressupostos fundamentais: informações de entrada e parâmetros em modelos caixa preta que não são diretamente observáveis. A adoção desses sistemas em domínios de aplicação sensíveis e de larga escala envolve várias questões éticas. Uma maneira de promover esses requisitos éticos é melhorar a explicabilidade desses modelos. No entanto, a explicabilidade pode ter objetivos e conteúdos diferentes de acordo com o público-alvo (desenvolvedores, especialistas de domínio e usuários finais). Algumas explicações nem sempre representam os requisitos dos usuários finais, pois desenvolvedores e usuários não compartilham o mesmo significado social. Este trabalho propõe um modelo conceitual, baseado na Engenharia Semiótica, que explora o problema da explicação como um processo comunicativo, no qual *designers* e usuários trabalham juntos em requisitos de explicações. O Modelo para Refletir sobre Inteligência Artificial Explicável (MoReXAI) é baseado em uma conversa estruturada, que promove a reflexão sobre temas como privacidade, justiça, responsabilidade, equidade e explicabilidade, visando ajudar os usuários finais a entender como os sistemas funcionam e apoiar a explicação. O modelo pode funcionar como uma ferramenta epistêmica, dadas as reflexões levantadas nas conversas relacionadas aos temas de princípios éticos, que auxiliaram no processo de levantamento de requisitos importantes para o desenho da explicação.

Palavras-chave: engenharia semiótica; ética; explicações; inteligência artificial.

ABSTRACT

The interest in systems that use machine learning has been growing in recent years. Some algorithms implemented in these intelligent systems hide their fundamental assumptions, input information and parameters in black box models that are not directly observable. The adoption of these systems in sensitive and large-scale application domains involves several ethical issues. One way to promote these ethics requirements is to improve the explainability of these models. However, explainability may have different goals and content according to the intended audience (developers, domain experts, and end-users. Some explanations does not always represent the requirements of the end-users, because developers and users do not share the same social meaning system, making it difficult to build more effective explanations. This paper proposes a conceptual model, based on Semiotic Engineering, which explores the problem of explanation as a communicative process, in which designers and users work together on requirements on explanations. A Model to Reason about the eXplanation design in Artificial Intelligence Systems (MoReXAI) is based on a structured conversation, with promotes reflection on subjects such as Privacy, Fairness, Accountability, Equity and Explainability, aiming to help end-users understand how the systems work and supporting the explanation design system. The model can work as an epistemic tool, given the reflections raised in the conversations related to the topics of ethical principles, which helped in the process of raising important requirements for the design of the explanation.

Keywords: semiotic engineering; ethics; explanations; Artificial intelligence.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, tem crescido bastante o interesse por sistemas que utilizam aprendizado de máquina (*Machine Learning* - ML). O uso desses sistemas, se justifica pelo fato de processarem um grande volume de dados, e fazer predições e tomar decisões a respeito de determinados contextos, com base nesses conjuntos de dados (*datasets*) (MOLNAR, 2020). A forma como esses sistemas impactam a vida das pessoas tem levantado grandes discussões éticas, pois esses algoritmos preditivos e os processos de decisão baseados nesses modelos matemáticos aprendem com dados do passado, os quais nem sempre é desejável repeti-los no futuro. Um clássico exemplo é o viés identificado no sistema de recrutamento da Amazon, que foi descartado de uso devido ao viés de gênero detectado, em função da base de dados utilizada nos últimos dez anos ter majoritariamente homens na área de tecnologia (DASTIN, 2018). Há também os casos em que alguns modelos escondem os pressupostos fundamentais, informações de entrada e parâmetros em modelos caixa preta, que não são diretamente observáveis, sendo necessário usar técnicas específicas para melhorar a interpretação das saídas geradas por eles (O'NEIL, 2016). Além disso, esses modelos são facilmente escaláveis e muitas vezes são implementados em populações ou situações para as quais não foram concebidos (SOUZA, 2018).

Um outro exemplo dos impactos negativos que afetam princípios éticos na utilização desses sistemas em larga escala é o uso do COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) (ANGWIN *et al.*, 2016). Esse sistema utiliza aprendizado de máquina para avaliar infratores nos Estados Unidos. Desenvolvido em 1998, teve o componente de avaliação de reincidência colocado em funcionamento a partir de 2000. Este *software* prevê o risco de um réu reincidir em um crime ou cometer uma contravenção dentro de dois anos. Essa avaliação se dá a partir de 137 caraterísticas dos detentos, obtidas a partir de perguntas, e seus antecedentes criminais (DRESSEL; FARID, 2018). Um estudo realizado para analisar a eficácia do COMPAS em mais de 7.000 indivíduos presos no condado de Broward, Flórida, entre 2013 e 2014, indicou que as predições fornecidas pelo sistema não eram confiáveis e tendiam a ser racistas, muito embora a raça não fosse considerada como característica de entrada no modelo (LARSON *et al.*, 2016). A acurácia do COMPAS para a predição de reincidência dos réus brancos era de 67,0%, ligeiramente superior a acurácia de 63,8% para réus negros. No entanto, os resultados afetaram réus negros e brancos de forma diferente: os réus negros que não reincidiram foram incorretamente preditos para reincidir em uma taxa de 44,9%, quase duas vezes mais alta que seus equivalentes brancos com 23,5%; e os réus brancos que reincidiram

foram incorretamente previstos para não reincidir em uma taxa de 47,7%, quase duas vezes maior que seus colegas negros com 28,0% (FLORES *et al.*, 2016).

A busca por Inteligência Artificial (IA) ética e responsável, para um melhor aproveitamento de todo o potencial que essas tecnologias podem desempenhar na sociedade, tem direcionado olhares de vários órgãos públicos e instituições privadas. A Comissão Europeia, por exemplo, definiu orientações éticas para IA (COMMISSION, 2020), na qual constam os direitos fundamentais e os princípios éticos para IA, bem como diretrizes para serem seguidas na busca por essa IA confiável. A explicabilidade é um dos requisitos para a transparência algorítmica, bem como um dos princípios éticos a serem alcançados no contexto desse documento. A explicabilidade diz respeito à capacidade de explicar tanto os processos técnicos de um sistema de IA, como as decisões humanas com eles relacionadas. A diretriz define que as decisões tomadas por um sistema de IA possam ser compreendidas e rastreadas por humanos e deve ser oportuna e adaptada ao nível de especialização da parte interessada (leigo, regulador ou auditor).

No Brasil, o direito a explicação foi instituído na LGPD (Lei Geral de Proteção de Dados) (BRASIL, 2018), aprovada em 2018, tomando como base a GDPR (*General Data Protection Regulation*) (EUROPÉIA,), da União Europeia, aprovada em 2016. A LGPD prevê o direito à explicação no caso de decisões totalmente automatizadas que possam ter um impacto na vida do titular dos dados. A explicação deve incluir não somente informações sobre os dados pessoais que serviram de entrada para o algoritmo, mas também sobre a lógica por trás de tais decisões. O direito também se aplica quando houver o tratamento de dados anonimizados, principalmente quando utilizado na formação de perfis comportamentais de pessoas identificadas (MONTEIRO, 2018).

A necessidade de sistemas inteligentes mais transparentes para atender legislação vigente, aumentar a confiança dos usuários e atender aos aspectos éticos, tem aumentado as pesquisas em um campo da IA chamada *eXplainable AI* (XAI). A Inteligência Artificial Explicável, refere-se a técnicas para melhorar a interpretabilidade ou explicabilidade de modelos de aprendizagem de máquina. Um sistema interpretável é aquele cujas operações são compreensíveis para humanos, seja por meio da inspeção do sistema, seja por meio de alguma explicação produzida durante o seu funcionamento (BIRAN; COTTON, 2017). A interpretabilidade tem forte relação com a transparência algorítmica. Modelos opacos como as redes neurais profundas necessitam de técnicas *post-hoc* para melhorar a compreensão do seu funcionamento interno.

O processo de desenvolvimento de sistemas de aprendizagem de máquina geralmente

envolve uma equipe multidisciplinar. Assim, para a construção de explicações mais eficazes é necessário um alinhamento por parte de todos os *stakholders*, sobre o artefato que pretendem criar, levantando hipóteses sobre o problema, experimentando diferentes possibilidades de solução e avaliando os resultados (BRANDÃO *et al.*, 2019). No entanto, existe uma dificuldade na definição dos objetivos de uma explicação, por parte dos *stakeholders*, tendo em vista a variedade do público-alvo (especialistas em IA, especialistas de domínio e usuários finais), a quem as explicações se destinam. Cada tipo de usuário necessita de níveis diferentes de detalhamento das explicações e objetivos diferentes quanto a explicabilidade dos modelos de IA (BRENNEN, 2020).

Embora exista um crescente campo de estudo que inclui usuários dentro desse processo investigativo de busca por melhoria da explicabilidade e das explicações dos modelos de *Machine Learning* (ML) (SOUZA, 2018), estas pesquisas e práticas, geralmente trazem a visão dos desenvolvedores sobre o que constitui uma "boa" explicação, desconsiderando os objetivos dos usuários nesses sistemas, o que pode levar ao fracasso (MILLER *et al.*, 2017). Miller (2019) argumenta que um dos caminhos para construir boas explicações é agregar conhecimentos do campo das ciências sociais, que já possuem um corpo de estudo bastante consolidado na área de explicações, trazendo várias abordagens filosóficas e psicológicas para endereçar esse problema. Para Miller (2019) as explicações são sociais, são uma transferência de conhecimento, apresentada como parte de uma conversa ou interação e, portanto, são apresentadas em relação às crenças de quem explica sobre as crenças de quem recebe a explicação. Uma boa explicação não está limitada apenas à melhoria da explicabilidade como uma forma de buscar as causas para determinadas predições, mas deve considerar que uma explicação é uma interação entre dois papéis: quem explica (emissor) e quem recebe a explicação (receptor), e que essa transmissão (comunicação) é um processo social. Assim, uma forte compreensão de como as pessoas definem, geram, selecionam, avaliam e apresentam explicações parece quase essencial (MILLER, 2019). Carbonera *et al.* (2018) propõem uso da Teoria da Engenharia Semiótica (SOUZA *et al.*, 2005), para explorar o problema de explicações como um processo comunicativo.

A Engenharia Semiótica (EngSem), é uma teoria fundamentada na significação e na comunicação e nos permite entender os fenômenos envolvidos no *design*, uso e avaliação de um sistema interativo. O *designer* comunica ao usuário, através da interface do sistema a quem ela se destina, que problemas ela pode resolver, e como interagir com ele (SOUZA *et al.*, 2005). Nesse sentido, o *designer* tem papel ativo no *design* da interação, pois é ele quem define quais

signos e sistema de significação serão utilizados para compor a interface, e atingir o objetivo de comunicar a sua mensagem.

Uma explicação, nesse contexto, é uma mensagem do *designer* para o usuário sobre o modelo produzido por ele. Para que essa comunicação seja efetiva este precisa, minimamente, compartilhar do mesmo sistema de significação e incluir nesta explicação o significado social dos sistemas que estão sendo projetados, levando em consideração objetivos, contextos de uso, aspectos culturais e éticos do público-alvo. Como a cultura influencia a comunicação humana, se faz necessário uma maior compreensão de quem são os envolvidos nesse processo comunicativo, suas atividades, experiências, valores e expectativas, para permitir uma melhor transmissão da metamensagem, a partir da interação com o sistema (LEITE, 1998). Muitas vezes o significado social não é pensado pelos desenvolvedores de IA, sendo necessário um processo de mediação por parte de um outro campo de estudo como a Interação Humano-Computador (IHC). *"Refletir a respeito do significado social dos sistemas de IA, é incluir um pensamento pragmático no processo de desenvolvimento desses sistemas e pensar em como eles podem afetar os usuários finais de forma direta ou indireta"* (BRANDÃO *et al.*, 2019).

O objetivo deste trabalho é apoiar *designers* na elicitação de requisitos de explicações a usuários finais no contexto de inteligência artificial explicável. Mais especificamente objetivamos explorar o problema de XAI como um processo comunicativo fundamentando-se na Engenharia Semiótica. Para isso, iremos investigar como especialistas em IA percebem explicações a usuários finais no contexto de desenvolvimento de modelos de ML, e explorar como os modelos e métodos da Engenharia Semiótica podem apoiar o processo de *design* e avaliação de explicações. Ao final, propomos um modelo conceitual, no qual *designers* e usuários trabalham juntos em requisitos de explicações. O Modelo para Refletir sobre Inteligência Artificial Explicável (MoReXAI) é baseado em conversas estruturadas, em que *stakeholders* (usuário, *designer* de interação, cientista de dados, programador, gerente de projeto, cliente etc.) respondem a perguntas pré-definidas e relacionadas a todo o processo de desenvolvimento de modelos de aprendizagem de máquina. As perguntas contidas no modelo possuem relação direta com princípios éticos como Privacidade, Justiça, Responsabilidade, Equidade e Explicabilidade. Estas conversas têm por objetivos extrair requisitos de explicações a usuários finais, levando em consideração o sistema de significação trazido durante as conversas, melhorando assim o processo de comunicação entre *designers* e usuários através da explicação produzida. O MoReXAI deverá funcionar como ferramenta epistêmica para o *design* de explicações, levando *stakeholders*

a refletirem sobre o produto que estão desenvolvendo e seu significado social no contexto do usuário final.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo apresentaremos os principais conceitos necessários para o desenvolvimento desta pesquisa. Na Seção 2.1 trazemos as definições relacionadas a explicação e explicabilidade na área de IA e sua relação com a IHC. Na Seção 2.2 são apresentados os conceitos que envolvem a Engenharia Semiótica (EngSem), teoria que será utilizada para fundamentar um modelo de discussão sobre Inteligência Artificial Explicável (XAI).

2.1 Inteligência Artificial Explicável (XAI)

As pesquisas na área de explicações em IA têm crescido, impulsionadas pela busca por maior entendimento do raciocínio subjacente em algoritmos caixa preta, e uma maior cobrança da sociedade por responsabilidade algorítmica. A área de explicações em IA não é recente e podem ser dividas em três gerações (MUELLER *et al.*, 2019). As primeiras pesquisas em explicações sugiram na década de 70, com os sistemas especialistas, e funcionavam criando expressões das regras lógicas e probabilísticas, descrevendo os objetivos e as etapas subjacentes usadas para fazer o diagnóstico ou tomar uma decisão. Assim, uma explicação era um *trace* das regras de como o problema era resolvido, uma tradução simples dos códigos utilizados para descrições de texto, apresentando pouca semelhança com a linguagem natural humana (SWARTOUT *et al.*, 1991). Apesar de ser um grande avanço na época, algumas regras para fazer as inferências e a tomada de decisão muitas vezes não faziam sentido para os usuários finais. Eram explicadas as regras que fizeram chegar ao resultado, mas o motivo pelo qual a regra estava correta não era representado (MUELLER *et al.*, 2019).

Uma explicação não deveria apenas ler o que aconteceu dentro do programa, ela deveria ser como uma interação estruturada que leva em consideração os objetivos, os antecedentes e o contexto do usuário (MOORE; SWARTOUT, 1988). Assim, as explicações deveriam ter estratégias flexíveis de diálogos de forma a se adaptarem às necessidades dos usuários individualmente, bem como, alternativas para produzir respostas que pudessem fornecer uma explicação mais detalhada quando os usuários não estivessem satisfeitos com a primeira explicação dada, e deveriam ser capazes de interpretar perguntas, levando em consideração, no contexto do diálogo, as explicações ou solicitações anteriores que não foram compreendidas (SWARTOUT, 1983).

A segunda geração dos sistemas de explicações aproveitou toda a base de conhecimento, regras e explicações utilizadas nos primeiros sistemas especialistas, e passaram a buscar

novas formas de gerar explicações, dessa vez com foco no reuso e melhoria do processo de desenvolvimento. Assim, os Sistemas Baseados em Conhecimento focavam em fazer explicações sensíveis ao contexto, pois incluíam conhecimento sobre a pessoa, sobre a história, sobre os objetivos do usuário e sobre o domínio. A ideia principal é que estes sistemas contavam com uma base de conhecimentos taxonômicos específicos de domínio e uma base estratégica de conhecimento separada (SWARTOUT; MOORE, 1993), e mais tarde, sendo acrescentada uma outra camada de comunicação (BARZILAY *et al.*, 1998). A separação da camada de comunicação do restante do sistema foi projetada para permitir que um especialista em comunicação crie soluções independentes do sistema e do domínio específicos.

A terceira geração dos sistemas de explicação, iniciou nos estudos de interpretabilidade de algoritmos de aprendizagem de máquina e seguem até os dias atuais, impulsionados pela aplicação desses modelos em larga escala em diversas áreas, o que implica na busca por mais transparência algorítmica e responsabilidade (MILLER, 2019). As pesquisas nos sistemas de explicações atuais, concentram-se, em sua maioria, na melhoria da explicabilidade dos algoritmos caixa preta, como a visualização, comparação e entendimento do raciocínio subjacentes em redes neurais, e explicações a usuários finais.

Para além da terceira geração, atualmente as pesquisas na área de sistemas de explicação, trazem direcionamentos na busca por explicações globais, que forneçam informações sobre como os modelos funcionam, ao invés de uma determinação específica para um caso específico (explicação local). Busca-se também por explicações que ajudem os usuários a desenvolver modelos mentais, saindo de formalismos como diagramas de algoritmos, diagramas mostrando traces de regras e cadeias de raciocínio, árvores de decisão, gráficos de barras, consultas mapeadas em gráficos de ontologia. Além desse conjunto de formatos, as explicações também assumem outras representações para atender usuários finais que têm outros objetivos nesses sistemas, como a utilização de redes neurais recorrentes (RNNs) no desenvolvimento de habilidades de linguagem para criação de *chatbots* e sistemas de respostas a perguntas (MILLER, 2019).

2.1.1 Conceitos e definições em Inteligência Artificial Explicável

As máquinas superam os humanos em muitas tarefas, e mesmo que não seja tão boa quanto um ser humano em alguma tarefa, devido a velocidade, reprodutibilidade e escala, ainda assim se torna vantajoso delegar algumas tarefas a elas. Um modelo de aprendizado de máquina

implementado pode completar uma tarefa muito mais rápido do que humanos, entregar resultados consistentes e confiáveis e ser copiado infinitamente. A replicação de um modelo de aprendizado de máquina em outra máquina é rápido e barato (MOLNAR, 2020). Aprendizagem de máquina ou ML, é um campo de estudo que dá aos computadores a habilidade de aprender com dados, sem ser explicitamente programado (GÉRON, 2019). Ela estuda métodos que computadores usam para melhorar predições ou comportamentos baseados em dados (MOLNAR, 2020).

Existem quatro principais categorias de aprendizado de máquina (GÉRON, 2019), as quais são indicadas a seguir:

- **Aprendizado supervisionado**: faz uso de conjuntos de dados rotulados para treinar algoritmos. Esse conjunto de dados de treinamento inclui entradas e saídas corretas, que permitem que o modelo aprenda ao longo do tempo, a partir do ajuste de parâmetros. Existem dois tipos principais de aprendizagem supervisionada: classificação e regressão. Na classificação o algoritmo é treinado para classificar os dados de entrada em variáveis discretas ou categóricas, e, na regressão, o algoritmo é treinado para prever uma saída a partir de uma faixa contínua de valores possíveis. São exemplos de algoritmos de aprendizado supervisionado: KNN (k-Nearest Neighbours), Regressão Linear, Regressão Logística, Máquina de Vetores de Suporte (SVM), Árvores de Decisão e Florestas Aleatórias e Redes Neurais (GÉRON, 2019).
- **Aprendizado não supervisionado**: os dados não são rotulados. O algoritmo não recebe durante o treinamento os resultados esperados, devendo descobrir por si só, por meio da exploração dos dados, os possíveis relacionamentos entre eles. O processo de aprendizado busca identificar regularidades entre os dados a fim de agrupá-los ou organizá-los em função das similaridades que apresentam entre si (ESCOVEDO; KOSHIYAMA, 2020).
- **Aprendizado semisupervisionado**: os algoritmos podem lidar com uma grande quantidade de dados não rotulados e dados parcialmente rotulados. Este aprendizado geralmente é uma combinação de algoritmos supervisionados e não supervisionados (GÉRON, 2019).
- **Aprendizado por reforço**: um agente pode observar o ambiente, selecionar e executar ações e obter recompensas positivas ou penalidades. Ele deve aprender por si só qual a melhor estratégia (política) para obter o maior número de recompensas ao longo do tempo. As políticas definem qual melhor ação o agente deve escolher quando está em determinada situação (GÉRON, 2019).

Apesar dos grandes benefícios dos modelos de aprendizagem de máquina, alguns

destes modelos possuem estruturas cada vez mais complexas, que tornam difícil interpretar o seu funcionamento e obter *insights* sobre os dados e a tarefa que eles resolvem. Modelos que utilizam florestas aleatórias (*Ramdom Forest*), por exemplo, são formadas por centenas de árvores de decisões, o que torna a tarefa de verificação de seu funcionamento interno muito difícil e complexa. Os modelos de melhor desempenho geralmente são combinações de vários modelos (*essembles*) que não podem ser interpretados, mesmo que cada modelo possa ser interpretado individualmente. Existe uma relação entre desempenho e opacidade, na qual, se você se concentrar apenas no desempenho, obterá automaticamente modelos cada vez mais opacos (ARRIETA *et al.*, 2020).

A IA explicável (XAI) refere-se a um conjunto de processos e técnicas de IA que permitem que usuários possam entender e confiar nos resultados e saídas geradas por um modelo de aprendizagem de máquina. Como campo de pesquisa, ela investiga a interpretabilidade (ou explicabilidade) e as formas de fornecer explicações de modelos e decisões algorítmicas. Em muitos trabalhos os termos "interpretabilidade"e "explicabilidade"têm o mesmo significado. No entanto, em nosso trabalho, utilizaremos a definição de Arrieta *et al.* (2020), na qual a interpretabilidade pode ser entendida como uma característica passiva de um modelo e que tem relação direta com o quanto ele é transparente e compreensível, por si só, para um humano (ARRIETA *et al.*, 2020). Já a explicabilidade é uma característica ativa do modelo e tem relação com ação ou procedimentos realizado com a intenção de esclarecer ou detalhar suas funções internas. Portanto, a explicabilidade refere-se ao contexto de explicações, e às inúmeras maneiras de trocar informações sobre como um modelo funciona ou o raciocínio subjacente utilizado para sua tomada de decisão, levando em consideração o público ao qual essa explicação é direcionada (ARRIETA *et al.*, 2020).

Dizemos que um modelo de aprendizado de máquina é considerado interpretável ou transparente se puder ser simulado (**simulatabilidade**), ou seja, puder ter todas as suas partes contempladas (parâmetros, entradas e saídas) por um ser humano, ser decomposto (**decomposabilidade ou inteligibilidade**) de forma que parte do modelo (entrada, raciocínio e saída) possa ser explicada separadamente e possuir **transparência algorítmica**, ou seja, o usuário seja capaz de entender como ele atuará em todas as situações que ele possa enfrentar, tendo em vista a possibilidade de ser explorado por meio de análises e métodos matemáticos (LIPTON, 2018). Essas três propriedades definem o grau de interpretabilidade de um modelo. Alguns modelos como regressão linear, árvores de decisão, KNN (*K-Nearest Neighbors*), aprendizagem

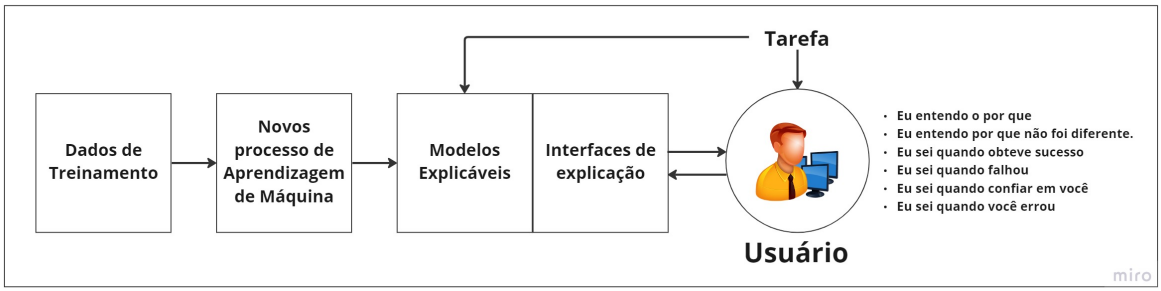
baseada em regras, modelos *bayseanos* e modelos generativos aditivos, são interpretáveis, e possuem algumas ou todas as propriedades citadas acima. Para a construção de boas explicações é necessário melhorar o processo de interpretabilidade de modelo.

Quando os modelos de aprendizagem de máquina possuem pouca interpretabilidade, dizemos que são opacos e necessitam aplicação de técnicas de interpretabilidade *post-hoc* para torná-los mais explicáveis. Várias técnicas *post-hoc* podem ser aplicadas como uma forma de melhorar a explicabilidade, por exemplo: formatos textuais ou símbolos que tentam retratar a lógica do algoritmo por meio de mapeamento semântico do modelo; explicações visuais que buscam representar o comportamento do modelo reduzindo a dimensionalidade e permitindo uma visualização simples interpretável para humanos; explicações locais que segmentam a solução fornecendo explicações para subespaços menos complexos, mas que são relevantes para todo o modelo; explicações que consideram a extração de exemplos relacionados ao resultado gerado por um determinado modelo, possibilitando uma melhor compreensão do próprio modelo; explicações por simplificação, nas quais é construído um modelo novo simplificado, com base no modelo a ser explicado, porém reduzindo a sua complexidade e mantendo desempenho semelhante ao original; explicação por relevância das *features*, na qual busca esclarecer o funcionamento interno de um modelo calculando uma pontuação de relevância para suas variáveis. Essas pontuações quantificam a importância dessa *feature* para a saída do modelo. Modelos de IA como florestas randômicas, SVM (*Suport Vector Machine*), redes neurais convolucionais, redes neurais recorrentes, entre outros, utilizam técnicas de interpretabilidade *post-hoc*.

O processo de melhoria da explicabilidade dos modelos de aprendizagem de máquina passa por pesquisas no campo das técnicas de interpretabilidade, desenvolvendo modelos mais explicáveis, como também pesquisas que envolvam a construção de interfaces de explicações que possam atender os objetivos dos usuários dos sistemas que se utilizam desses modelos (Figura 1).

O processo de desenvolvimento de modelos de aprendizagem de máquina implica em atividades em várias etapas, chamadas de pipeline. Toreini *et al.* (2020b) divide esse processo em duas grandes etapas (Figura 2): uma centrada nos dados (coleta dos dados, preparação dos dados e engenharia de *features*) e outra centrada no modelo (treino, teste e inferências). Primeiro, coleta-se os dados de uma fonte, em seguida esses dados passam por métodos de pré-processamento para depois serem extraídas determinadas características (*features*). Quando

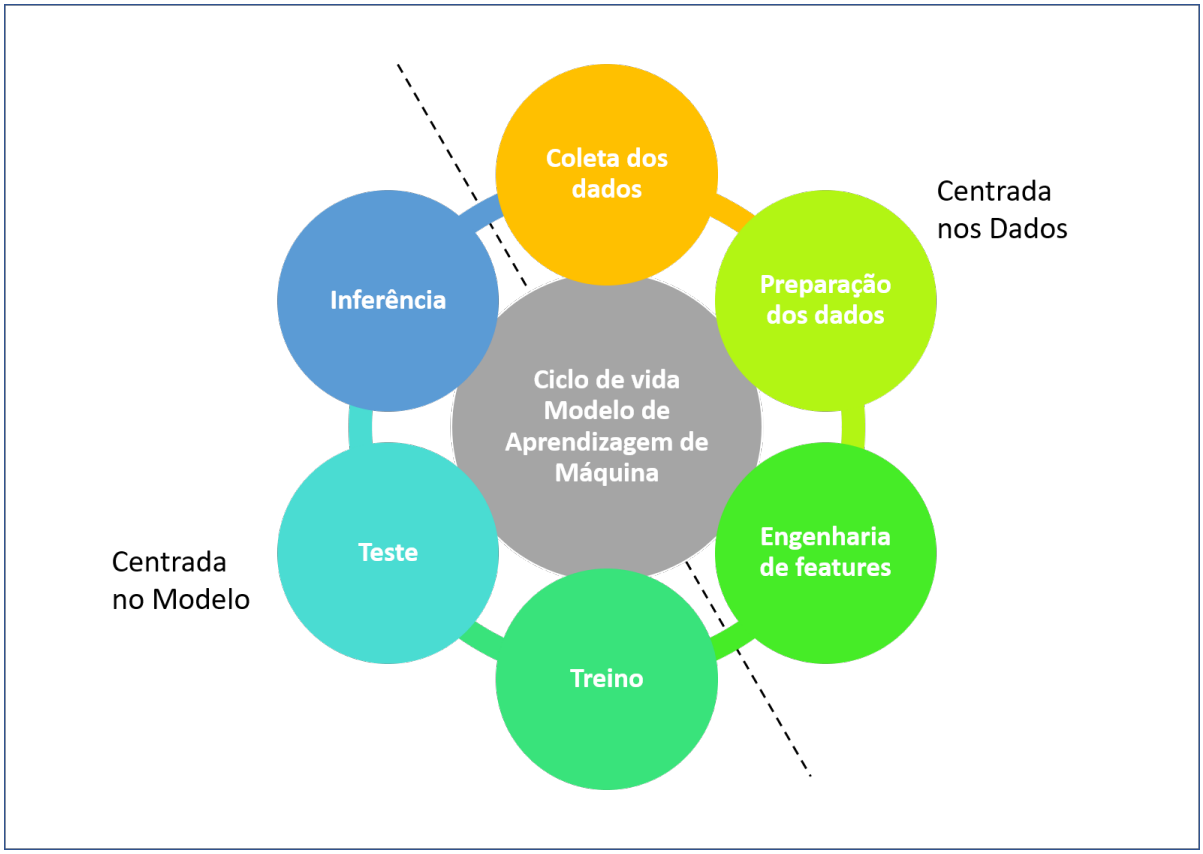
Figura 1 – Conceito de XAI



Fonte: Adaptado de Darpa (2016).

as *features* estão preparadas, são divididas em pelo menos dois grupos, treinamento e teste, sendo o primeiro responsável pelo aprendizado do modelo e o segundo por realizar avaliação do modelo e calibrar os parâmetros para melhoria da métrica escolhida de avaliação. Quando o modelo passa no estágio de verificação com um desempenho suficientemente bom, eles são aplicados no mundo real (TOREINI *et al.*, 2020a).

Figura 2 – Ciclo de desenvolvimento dos modelos de aprendizagem de máquina



Fonte: Adaptado de (TOREINI *et al.*, 2020a).

2.1.2 Design de explicações em IA

O processo de *design* de explicação, assim como o processo de *design* de interação e interface passa, primeiramente, pela definição dos objetivos e intenções que se quer alcançar com aquela explicação, tendo como foco o público-alvo ao qual ela se destina. Nesta seção abordaremos os objetivos, o público-alvo e os formatos de explicações, assim como o que caracteriza uma boa explicação e abordagens de explicações a usuários finais.

Os usuários, em geral, são os principais públicos da XAI, pois, na maioria das vezes, são eles que dependem de decisões, recomendações ou ações produzidas por esses modelos de ML e, portanto, precisam entender a razão para as decisões do sistema. O conteúdo das explicações e sua forma de apresentação são influenciados fortemente pelos usuários-alvo a qual elas se destinam. Mohseni (2019) classifica os usuários conforme apresentado a seguir:

- **Novatos em IA**: referem-se a usuários finais que usam produtos da IA no dia a dia, mas não possuem (ou possuem pouca) experiência em sistemas de aprendizado de máquina. Usuários que utilizam sistemas de recomendação, ou recebem diagnóstico médico, ou fazem solicitações de empréstimos, buscam explicações sobre as inferências geradas pelos modelos.
- **Especialistas em dados ou especialistas de domínio**: são usuários que utilizam o aprendizado de máquina para análise, tomada de decisão ou pesquisa. Esse grupo de usuários costuma utilizar ferramentas inteligentes de análise de dados ou sistemas de análise visual para obter *insights* dos dados.
- **Especialistas em IA**: cientistas e engenheiros que desenvolvem algoritmos de aprendizagem de máquina e técnicas de interpretabilidade. Essas técnicas podem estar relacionadas à interpretação intrínseca dos modelos ou às explicações (interpretabilidade) *post-hoc*.

Mohseni (2019) também especifica os objetivos dos usuários quanto a necessidade das explicações, conforme descrito a seguir:

- **Transparência algorítmica:** a melhoria na transparência ajuda os usuários a entender como o modelo funciona, permitindo obter novos *insights* sobre como modelos aprendem padrões de dados e melhorando a experiência do usuários e a interatividade com o sistema.
- **Confiabilidade:** a confiabilidade tem relação com o modelo agir como pretendido e tem forte relação com a robustez e a estabilidade. Esse objetivo não está associado somente ao desempenho dos algoritmos, mas também ao sentido de gerar saídas confiáveis em cenários reais. Por exemplo, um modelo preditivo, que prevê taxas de criminalidade em

uma determinada localidade, pode ter uma alta acurácia, porém não ser confiável, no sentido de conter viés e fazer inferências discriminatórias. A confiabilidade gera confiança dos usuários no sistema e em suas predições.

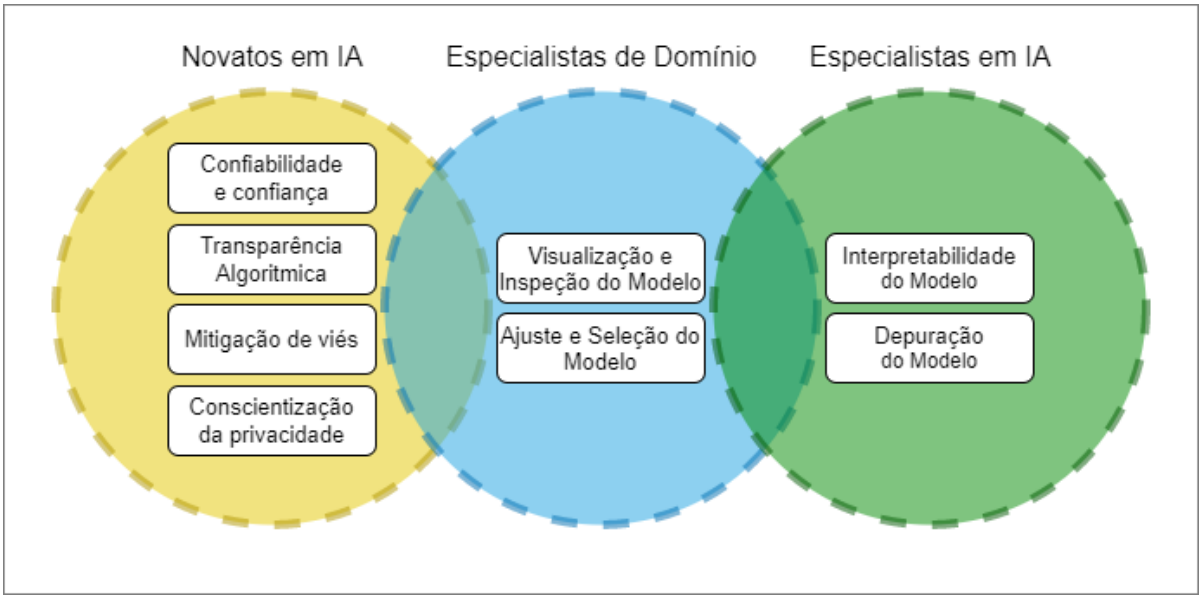
- **Confiança:** esse objetivo diz respeito à confiança dos usuários no sistema e em suas predições a partir das explicações geradas por eles. Aplicações como sistemas de recomendação, sistemas autônomos, e sistemas de tomada de decisão crítica são exemplos de sistemas que visam melhorar a confiança do usuário no seu uso.
- **Redução de viés:** as explicações podem ajudar os usuários finais a inspecionar se os sistemas são tendenciosos ou se possuem algum viés discriminatório. A partir das explicações do comportamento dos algoritmos, os usuários podem perceber dados de treinamento tendenciosos ou utilização de *features* em modelos de ML que podem resultar em viés algorítmico.
- **Conscientização sobre privacidade:** a explicação ajuda os usuários a avaliarem a privacidade dos seus dados, revelando quais desses dados estão sendo usados na tomada de decisão algorítmica.
- **Visualização e inspeção do modelo:** a explicabilidade ajuda os especialistas de domínio a identificar e analisar falhas nos modelos e sistemas de aprendizado de máquina. Nesse sentido, buscar maior transparência do modelo por meio de técnicas de visualização e interação, melhorando a análise de dados.
- **Ajuste e seleção do modelo:** especialistas em dados necessitam, muitas vezes, ajustar os parâmetros dos modelos de forma visual e interativa, para analisar como os modelos se comportam. Essa possibilidade permite que os especialistas comparem vários modelos e selecionem o modelo certo para os dados-alvo.
- **Interpretabilidade dos modelos:** é a meta principal do especialista em IA e permite obter novos *insights* sobre como modelos de aprendizagem profunda aprendem padrões de dados.
- **Depuração do modelo:** busca, a partir de técnicas de interpretabilidade, melhorar a arquitetura do modelo e o processo de treinamento.

Esses objetivos não possuem domínio exclusivo para cada tipo de usuário final e em alguns momentos podem ser encontradas interseções entre os objetivos da IA explicável e os tipos de usuários, como apresentado na Figura 3. Nossa proposta se concentra no *design* de explicação sob a perspectiva dos novatos em IA e especialistas de domínio, aos quais iremos nos

referir como usuários finais, excluindo assim desse escopo os especialistas em IA.

A partir do contexto dos objetivos dos usuários, as explicações podem ser projetadas usando uma variedade de formatos para atender diferentes grupos. Explicações visuais usam elementos visuais para descrever o raciocínio por trás dos modelos de aprendizado de máquina, como mapas de atenção e mapas de calor. Explicações verbais descrevem o modelo ou raciocínio de máquina com palavras, frases ou linguagem natural. As explicações verbais são muito utilizadas no contexto de sistemas de recomendação. Portanto, as interfaces de explicação podem fazer uso de várias modalidades (por exemplo elementos visuais, verbais e numéricos) de explicações para apoiar a compreensão do usuário. Um formato bastante comum de explicações utilizadas por especialista de domínio são as explicações analíticas, as quais permitem visualizar e explorar dados e representações dos modelos de aprendizado de máquina e dependem de métricas numéricas e visualizações de dados. Essas ferramentas de análise visual também permitem que os pesquisadores revisem as estruturas do modelo e as relações dos seus parâmetros. Visualizações de mapa de calor, gráficos e visualizações hierárquicas (árvores de decisão) são comumente usadas para visualizar explicações analíticas para algoritmos interpretáveis (ARRIETA *et al.*, 2020).

Figura 3 – Usuários x Objetivos de *design* da IA Explicável



Fonte: Adaptado de Mohseni (2019).

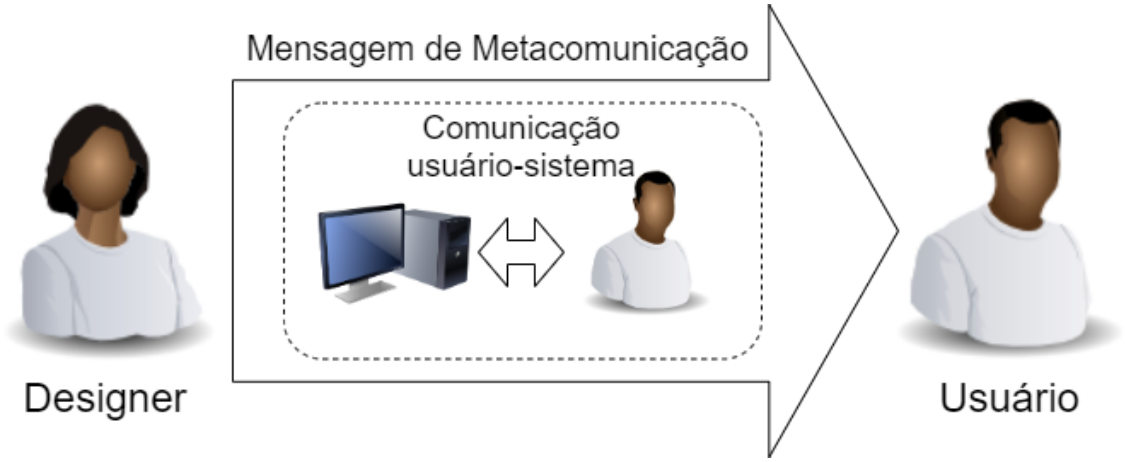
2.2 Engenharia Semiótica

A Teoria da *Engenharia Semiótica* (EngSem) é uma teoria explicativa de IHC que nos permite entender os fenômenos envolvidos no *design*, no uso e na avaliação de um sistema interativo (SOUZA *et al.*, 2005). Nesta teoria um *software* é um artefato intelectual, ou seja, é um produto criado a partir da interpretação de um projetista sobre um problema e sua concepção de solução, que é então apresentada em uma codificação linguística (PRATES; BARBOSA, 2020). Este artefato é descrito em alguma linguagem artificial e processada por um computador, com a qual o usuário vai interagir através da interface. Na EngSem, a interação é vista como uma conversa entre *designer* e usuário através da interface (preposto) no momento que o usuário faz uso dela (Figura 4). A interface comunica para o usuário, a visão do projetista, relacionada a quem ela se destina, que problemas ela pode resolver e como interagir com ela. O conteúdo dessa mensagem pode ser entendido como:

"Esta é a minha interpretação sobre quem você é, o que eu entendi que você quer ou precisa fazer, de que formas prefere fazê-lo e por quê. Eis, portanto, o sistema que consequentemente concebi para você, o qual você pode ou deve usar assim, a fim de realizar uma série de objetivos associados com esta (minha) visão."

Nesse sentido, a mensagem transmitida pela interface é indireta, pois é entendida pelo usuário a partir da sua interação, e unidirecional, tendo em vista que o usuário não pode dar continuidade àquela comunicação.

Figura 4 – Metacomunicação *designer*-usuário e comunicação usuário-sistema



Fonte: Figura adaptada de Barbosa e Silva (2010).

A EngSem é fundamentada na Semiótica, e sua ontologia compreende os processos de significação e comunicação, os interlocutores envolvidos nesse processo e o espaço de *design*

de IHC (SOUZA *et al.*, 2005).

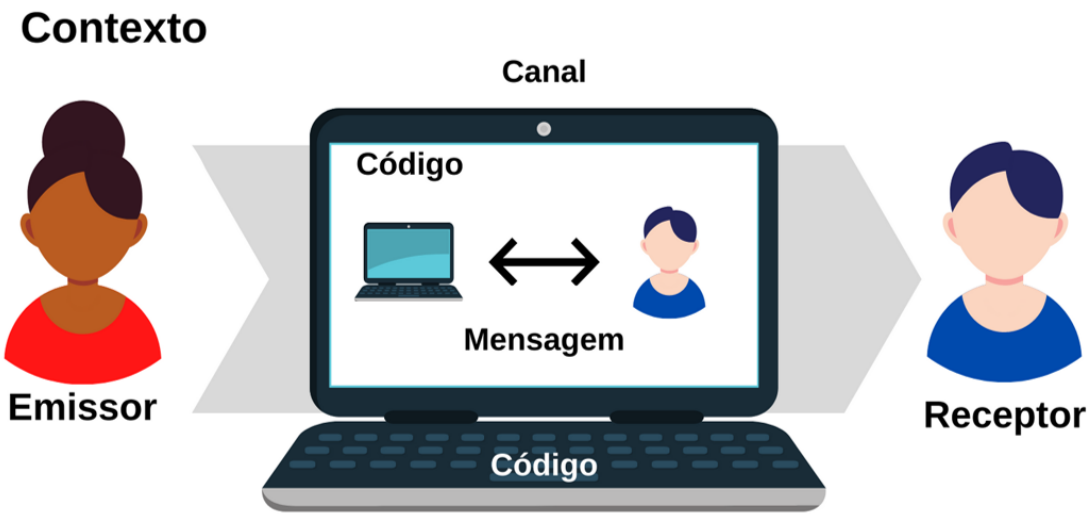
A significação é o processo através do qual, expressão e conteúdo, de signos, são estabelecidos com base em convenções sociais e culturais conhecidas das pessoas que vão utilizá-los. A esse processo de codificação dá-se o nome de sistema de significação, o qual, na computação, pode acontecer de forma artificial. Signo é tudo aquilo que significa algo para alguém (SANTAELLA; NOTH, 2021; PEIRCE, 1902). Nesse sentido ele relaciona três elementos: o *representamen*, que tem relação com a maneira com a qual o signo está representado; um objeto ou representante; e a ideia formada na cabeça da pessoa, referente àquele objeto, no qual se dá o nome de interpretante. Esse processo interpretativo que nos leva a associar cadeias de significados a um signo dá-se o nome de semiose (SANTAELLA; NOTH, 2021; PEIRCE, 1902). O processo de semiose é influenciado pelo contexto social e cultural, hábitos e experiência pessoal do intérprete.

A comunicação é o processo através do qual pessoas, utilizando-se de signos, de um ou mais sistemas de significação, produzem mensagens com o intuito de expressar determinados conteúdos. Nesse sentido, a cultura influencia a comunicação humana, tendo em vista o compartilhamento comum de signos e significados que convergem na forma de padrões de representação, utilizados na produção e na troca de mensagens entre indivíduos e grupos (SOUZA *et al.*, 2005). Portanto, como EngSem percebe o processo de *design* de interface como um ato comunicativo, no qual o *designer* deve tomar decisões sobre a solução que utilizará para compor a interface (definição de signos e sistemas de significação), se faz necessário uma maior compreensão de quem são os usuários, suas atividades, experiências, valores e expectativas, para permitir uma melhor transmissão da meta-mensagem a partir da interação com o sistema.

O espaço de *design* de IHC na EngSem é baseado no modelo de espaço de comunicação proposto por (JAKOBSON, 1960) e é estruturado em: contexto, emissor, receptor, mensagem, código e canal (Figura 5). Para projetar a meta-mensagem, o *designer* deve tomar decisões sobre cada elemento deste modelo. A Tabela 1, apresenta as perguntas que guiam o processo de *design* da meta-mensagem.

Portanto, o *designer* tem papel ativo na interação, tendo em vista que ele é interlocutor e deve ajudar os usuário a entender a meta-mensagem contida na interface. Para isso, deve refletir sobre os tipos de estratégias que deve utilizar, os signos que pode projetar na interface e as consequências que as limitações dos significados computacionais trazem para a interação (SOUZA; LEITÃO, 2009; BARBOSA; SILVA, 2010). Portanto, a construção da meta-mensagem

Figura 5 – Espaço de *design* de IHC da EngSem (de Souza *et al.*, 2001), com base no modelo de espaço de comunicação de Jakobson (1960)



Fonte: Figura adaptada de de Souza *et al.* (2001).

Tabela 1 – Perguntas que guiam o processo de *design* de interface na EngSem a partir do modelo de Jakobson (1960)

Elemento/Pergunta	Descrição
Quem é o emissor (*designer*)?	Que aspectos das limitações, motivações, crenças, e preferências do *designer* devem ser comunicados ao usuário para o benefício da meta-comunicação.
Quem é o receptor (usuários)?	Que aspectos das limitações, motivações, crenças, e preferências do usuário, tal como interpretado pelo *designer*, devem ser comunicados aos usuários reais para que eles assumam seu papel como interlocutores do sistema.
Qual o contexto da comunicação?	Que elementos do contexto de interação (psicológico, sociocultural, tecnológico etc.), devem ser processados pelo sistema.
Qual é o código da comunicação?	Que códigos computáveis podem ou devem ser utilizados para apoiar a metacomunicação eficiente, ou seja, qual deve ser a linguagem da interface.
Qual é o canal?	Quais canais de comunicação estão disponíveis para a metacomunicação *designer*-usuário e como eles podem ou devem ser utilizados.
Qual é a mensagem?	O que o *designer* deve contar aos usuários, e com que efeito, ou seja, qual é a intenção comunicativa do *designer*.

Fonte: elaborado pelo autor.
Nota: Tabela construída com base nas informações contidas em Barbosa e Silva (2010)

a ser enviada pelo *designer* é o próprio processo de *design* do sistema e acontece sob a perspectiva de reflexão durante a ação (SCHÖN, 1938; PRATES; SILVA, 2010), em que cada problema é visto como único, pois é caracterizado por elementos do contexto que definem o problema, e mutável, pois não existe uma única solução, e ela pode alcançar diferentes formas à medida que o *designer* e os usuários melhoram seu entendimento sobre o problema. Nesse sentido, segundo Schön e Bennett (1996), Prates e Silva (2010), na construção da interface, o *designer*

precisa refletir sobre o artefato que ele pretende criar, levantando hipóteses sobre o problema, experimentando diferentes possibilidades de solução e avaliando os resultados. Para isso ele utiliza ferramentas epistêmicas que permitem que ele reflita sobre questões relacionadas aos artefatos de metacomunicação e compare diferentes propostas de solução.

Diferentemente das abordagens de *design* centradas no usuário, que buscam produzir tecnologia com foco na usabilidade do sistema, considerando aspectos operacionais da interação, o *Design Centrado na Comunicação* (DCC) fundamentado na EngSem tem o foco na comunicação e nas estratégias adotadas pelo *designer* para comunicar aos usuários sua visão de *design* e dar-lhes melhores condições de entender e aprender sobre o sistema projetado.

O processo de *design* centrado na comunicação, proporciona uma compreensão compartilhada, da metacomunicação, por todos os membros da equipe de projeto. O compartilhamento da metacomunicação é uma oportunidade para que todos os envolvidos possam contribuir, a partir de sua perspectiva particular, a respeito do problema em questão. Essa compreensão compartilhada da metacomunicação deve ser registrada em todas as etapas do processo de *design*, para que possa ser revisada, evitando e corrigindo interpretações incompletas ou equivocadas por parte dos membros da equipe. Nesse contexto, o uso de ferramentas epistêmicas, durante as atividades de *design*, permitem a reflexão sobre questões relacionadas aos artefatos de metacomunicação através da comparação de diferentes propostas de solução. Uma ferramenta epistêmica ajuda aumentar o entendimento sobre o problema que está sendo resolvido, embora não dá uma resposta ou solução, como é o caso de diretrizes e regras (de Souza *et al.*, 2001).

A elaboração da metacomunicação é orientada por um conjunto de perguntas derivadas das dúvidas comuns dos usuários. São essas perguntas que geram insumos para cada atividade do processo de *design* e que foram propostas por Silveira *et al.* (2005) na arquitetura para construção de sistemas de ajuda *on-line*. Essa arquitetura é uma ferramenta epistêmica e tem o foco na comunicação direta do projetista para o usuário através do sistema de ajuda. Ela organiza a construção do sistema de ajuda através dos componentes: ajudas locais, módulo de ajuda geral, instruções diretas e mensagens de erro. Para cada tipo de elemento de interação disponível, a arquitetura oferece um conjunto de questões a serem respondidas pelo projetista para compor o sistema de ajuda e *templates* de respostas.

Além da arquitetura proposta por Silveira *et al.* (2005), várias outras ferramentas epistêmicas, fundamentadas na EngSem, foram propostas na literatura, como a Manas (BARBOSA *et al.*, 2007), que tem o objetivo de apoiar *designers* na representação e reflexão sobre

como os usuários podem ou devem se comunicar entre si através dos Sistemas Colaborativos, fornecendo indicadores acerca dos possíveis impactos sociais que esse modelo pode causar no grupo de usuários apoiados pelo sistema; a Modeling Language for Interaction as Conversation (MoLIC), uma linguagem para modelagem de interação, que permite representar a interação entre usuário e sistema como um conjunto de conversas na qual os usuários podem (ou devem) travar com o sistema (ou preposto de projetista) para atingir seus objetivos (BARBOSA; PAULA, 2003); e o Modelo para Descrever e Negociar Modificações em Sistemas *Web*, proposto por Sampaio (2010) com o objetivo de descrever e comunicar mudanças em sistemas de grupo na *Web* e negociando, a partir de um conjunto de códigos, o melhor *design* de interface para atender às demandas dos usuários.

3 TRABALHOS RELACIONADOS

Neste capítulo são apresentados alguns trabalhos que se relacionam com a proposta de criação de um modelo de comunicação entre *stakeholders* na construção IA Explicável.

3.1 *Design* fundamentado na EngSem e ética

O trabalho de Barbosa *et al.* (2021), é um exemplo de uso da Engenharia Semiótica para abordar questões éticas e de responsabilidade social na produção de artefatos digitais. Eles propuseram uma extensão do template de metacomunicação, artefato central usado na Engenharia Semiótica, para apoiar o *design* centrado no homem, de modo a abordar diretamente a responsabilidade moral e as questões éticas. Essa extensão funciona como uma ferramenta epistêmica e pode ser utilizada para criar e elaborar conhecimento relacionado a essas questões. A ideia é trazer pra dentro da construção da mensagem de metacomunicação a visão, não só do designer, mas de todos os envolvidos no processo de desenvolvimento da tecnologia digital. O principal ator deixa de ser "Eu"e passa a ser "Nós". As perguntas utilizadas para a construção da meta-mensagem agora são respondidas por todos os *stakeholders*, acrescentado, em cada etapa de *design*, questões que trazem uma reflexão ética a respeito do produto que desenvolvem e como esta tecnologia pode afetar os usuários. Este trabalho tem forte relação com a nossa proposta, por trazer a perspectiva de todos os *stakeholders* para o processo de *design*, através da discussão de questões éticas e de responsabilidade. No entanto, em nossa proposta, trazemos uma perspectiva mais focada dentro do processo de desenvolvimento de modelos de aprendizagem de máquina.

3.2 *Design* de explicação com foco em usuários finais

Existe trabalho recente sobre projeto de explicação com foco no usuário final, em Eiband *et al.* (2018a), eles propõem diretrizes para melhorar a transparência de algoritmos de IA. O conteúdo de uma explicação (o que explicar) é elicitado a partir das seguintes etapas: captura do modelo mental dos especialistas em IA e o que eles consideram ideal para os usuários; captura do modelo mental dos usuários; e síntese do modelo mental alvo, no qual são selecionados os principais componentes do modelo mental do especialista, que são mais relevantes para os usuários, e o nível de detalhe preferido por eles. Depois eles se concentram no formato de apresentação da explicação (como explicar), na qual primeiramente é realizada uma prototipagem iterativa, com base nas diretrizes de *design* corporativo e no fluxo de usuários

do sistema, explorando possíveis visualizações dos principais componentes do modelo mental alvo, em seguida é realizada uma avaliação de projeto, na qual as diferenças e combinações entre o modelo mental do usuário e o modelo mental alvo são investigadas para avaliar o *design* do protótipo. Os referidos autores argumentam que a transparência é um conceito difuso e multifacetado que abrange uma variedade de áreas de pesquisa e que deve ser projetada de maneira a beneficiar os usuários. Para isso, deve levar em consideração os requisitos dos usuários e de outras partes interessadas, porém levando em consideração que os usuários podem não estar interessados em todos os sistemas subjacentes de raciocínio e que acham algumas informações mais interessantes que outras. Utilizam técnicas de *design* participativo para capturar os modelos mentais dos usuários e especialistas em IA, utilizando abordagem centrada no usuário e Teoria da Engenharia Cognitiva. Em nosso modelo promovemos a reflexão sobre os aspectos sociais envolvidos nos sistemas de IA, com os *stakeholders*, utilizando abordagem centrada na comunicação e a Teoria da Engenharia Semiótica.

Em Mueller *et al.* (2021), os autores propõem o uso de princípios focados no ser humano para o *design*, teste e implementação de sistemas XAI de forma que sejam implementados algoritmos para atender a esse propósito. Eles elaboraram o *"Self-Explanation Scorecard"*, que pode ajudar os desenvolvedores a entender como eles podem capacitar os usuários, permitindo a autoexplicação. Apresentam também um conjunto de princípios de *design* centrados no usuário e empiricamente fundamentados, que podem orientar os desenvolvedores a criar sistemas explicáveis bem-sucedidos. Em nosso trabalho nós também usamos uma abordagem envolvendo usuários finais, entretanto, utilizamos o *design* focado na comunicação, no qual usuários e desenvolvedores discutem sobre explicações em um modelo de discussão previamente estruturada.

Em Lopes *et al.* (2021), os autores apresentam o processo de desenvolvimento, de uma ferramenta multiperspectiva, centrada no usuário, para a interpretabilidade de aprendizado de máquina chamada Explain-ML. A ferramenta foi projetada para implementar um fluxo de trabalho no qual o usuário pode realizar interativamente as etapas do ciclo de vida de um modelo de ML. Para cada projeto, ele pode criar várias execuções, alterando as configurações de definição e otimização de hiperparâmetros do modelo e gerando um conjunto de visualizações que transmitem aspectos relativos ao modelo, ao conjunto de dados de treinamento do modelo e também às informações especificas de instâncias. Essas visualizações atuam como explicações para o modelo projetado para disponibilizar diferentes perspectivas complementares entre si

(global, base de dados e local), as quais auxiliam o usuário na interpretação dos resultado do modelo. Além da definição da ferramenta, foi realizado um estudo qualitativo que permitiu analisar em profundidade a perspectiva e as percepções dos usuários sobre a ferramenta. A partir da análise sobre resultados obtidos na avaliação da experiência dos usuários com a Explain-ML, observaram potencial relevância para atender os princípios para projetos de interfaces de *Interative Machine Learning* (DUDLEY; KRISTENSSON, 2018), bem como consolidá-los.

Em Weitz *et al.* (2020) investigam se uma interface de usuário com um agente virtual tem um efeito positivo na confiabilidade percebida de um modelo de classificação baseado em Redes Neurais Artificiais. Para isso testaram se as modalidades de apresentação das informações (informação pura em forma de texto, voz ou presença visual), que foram escolhidas para a comunicação dos resultados de previsão do classificador e suas visualizações XAI, tiveram um impacto significativo na confiança dos usuários. Observaram que existe uma tendência de melhoria da confiança percebida do usuário de acordo com as modalidades de agente escolhido. Subindo de grupo sem agentes sobre grupos de texto e fala até grupo de agentes virtuais incorporados. Descobriram também que os usuários finais desejam explicações linguísticas adicionais e que desejam interagir com o agente fazendo perguntas. Observaram que a experiência dos usuários poderia se beneficiar de um *design* de interação XAI mais humano a partir do uso de agentes virtuais para atingir esse objetivo de *design*.

Em Zhou *et al.* (2020) investigam como o *design* explicável pode afetar a compreensão do usuário no campo da educação. Para isso tornam o algoritmo de inteligência artificial explicável acessível a não cientistas da computação na forma de uma experiência lúdica. Partem da intuição de que o *design* explicável a partir de uma experiência interativa é mais envolvente do que um ambiente de aprendizagem estático. Se utilizam de teorias educacionais, nas quais atividades de aprendizagem ativas aumentam o aprendizado dos alunos nas áreas de ciências, engenharia e matemática. Como resultado observam quem diferentes formações educacionais exigem diferentes abordagens para projetar um explicável. Nem todos os usuários estavam interessados em entender profundamente o algoritmo, sendo assim importante fornecer uma visão geral para todos os usuários, mas também uma opção de detalhes sob demanda. Os usuários serão engajados de forma diferente com os conteúdos explicáveis. Nem todos os usuários acham o aprendizado sobre algoritmos intrinsecamente agradável. Motivar o conteúdo explicável de várias perspectivas pode ajudar, assim como criar uma experiência explicável envolvente.

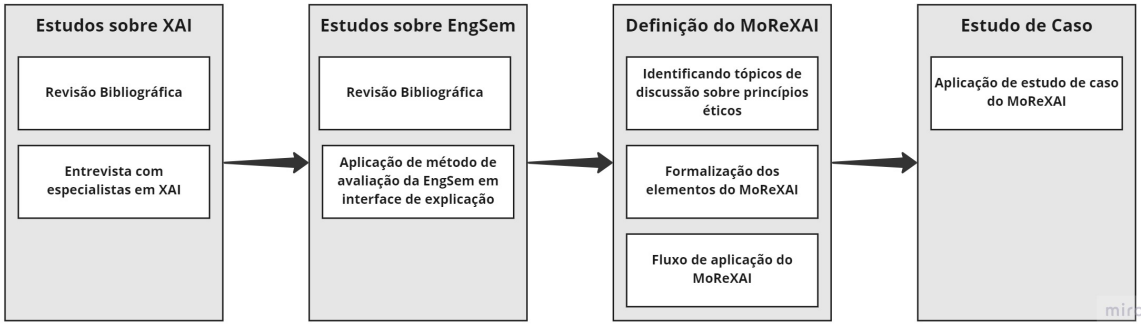
3.3 *Design* de explicação com foco em usuários finais e EngSem

Assim como nossa pesquisa, o trabalho de Ferreira e Monteiro (2021) é sobre XAI para usuários finais. Eles argumentam que é necessário discutir o XAI no início do processo de *design* do sistema de IA e com todas as partes interessadas. Com isso visam investigar como operacionalizar a discussão sobre cenários e oportunidades de XAI entre *designers* e desenvolvedores de IA e seus usuários finais. Para eles o *design* de explicações em IA tem várias dimensões que devem ser consideradas, por exemplo: "Quem"são as pessoas interessadas em explicações de IA (usuários finais, tomadores de decisão, usuários afetados, órgãos reguladores, construtores de sistemas de IA); as motivações e expectativas que cada uma das pessoas interessadas tem para as explicações ("Por quê"). Assim, eles tomam a Engenharia Semiótica como base teórica e a *SigniFYIng Message* ("quem", "o quê", "por quê", "como", "quando"e "onde") como ferramenta conceitual para estruturar as diferentes dimensões (usuários e desenvolvedores) que devem ser consideradas para a discussão dos cenários XAI. Esse trabalho se relaciona com o nosso por abordar o problema de *design* de explicações a partir das lentes da Engenharia Semiótica, reconstruindo a mensagem de metacomunicação a partir do *SigniFYIng Messagem*, no entanto não promovem a discussão entre os *stakholders* sobre questões éticas como propomos em nossa pesquisa.

4 METODOLOGIA

Neste capítulo apresentamos a metodologia para a construção do MoReXAI. Primeiramente realizamos pesquisa qualitativa com especialistas em IA para entender como eles percebem explicações em IA (Seção 4.1). Depois realizamos uma avaliação de explicações sob a ótica da Engenharia Semiótica e aprofundamos os estudos sobre os modelos e métodos desta teoria e como ela pode contribuir com o processo de explicações em IA (Seção 4.2). Definimos os elementos do Modelo para Refletir sobre Inteligencia Artificial Explicável (MoReXAI) e o seu fluxo de aplicação (Seção 4.3), por fim realizamos um estudo de caso no contexto de uma aplicação real (Seção 4.4).

Figura 6 – Metodologia para desenvolvimento do MoReXAI



Fonte: Elaborado pelo autor (2020).

4.1 Obtendo a visão dos especialistas sobre explicações em IA

Esta etapa da metodologia teve como objetivo explorar a área de IA, mais especificamente aprendizagem de máquina, seu funcionamento e como são geradas as explicações desses modelos de ML. Para isso realizamos revisão de literatura, na busca por um panorama geral da área e direcionar olhares para alguns estudos que seriam a base da pesquisa. Essas revisões trouxeram conceitos e artigos chave que foram importantes para o desenvolvimento do MoReXAI. A partir das leituras realizadas na revisão bibliográfica sobre XAI e os desafios de pesquisa para a construção de boas explicações a usuários finais, realizamos um estudo exploratório, utilizando um método de investigação qualitativo, com o objetivo de entender como especialistas em ML percebem explicações a usuários finais.

Realizamos entrevistas como método de investigação, pois, segundo Barbosa e Silva (2010), conversas diretas fornecem melhores dados e perspectivas, que muitas vezes passam

despercebidas na aplicação de métodos como questionários. No planejamento das entrevistas, optamos por usar perguntas abertas, sem restrições sobre o tempo ou tamanho da resposta. O roteiro da entrevista foi organizado de forma semiestruturada e composto por perguntas que seguem uma sequência lógica, de forma a explorar com mais profundidade as respostas fornecidas e mantendo o foco no objetivo. Para isso foram coletados todos os áudios das entrevistas para posterior análise e compilação dos resultados. A gravação dos áudios foi realizada utilizando o aplicativo gravador de voz, no modo entrevista, de um celular Samsung Galaxy S9, com Android na versão 9.

O roteiro da entrevista foi dividido em três etapas: *(i)* apresentação: em que foram apresentados os objetivos da entrevista e firmado o termo de consentimento; *(ii)* aquecimento: em foram feitas perguntas de fácil resposta para levantar dados demográficos, além de entender a proximidade dos entrevistados com ML; *(iii)* bloco principal: em que foram feitas perguntas que atendem o objetivo da pesquisa de forma mais especifica, contendo perguntas gerais a respeito de Aprendizado de Máquina e outras relacionadas a explicações a usuários. O conjunto de perguntas é apresentado na Tabela 2

Tabela 2 – Roteiro de perguntas da entrevista com especialistas em *Machine Learning*

(i)	**Apresentação do objetivo e termo de consentimento**
(ii)	**Perguntas de aquecimento** 1. Você já estudou ou desenvolveu algum modelo de *Machine Learning* e/ou IA? 2. Há quanto tempo você trabalha nessa área? 3. Como surgiu o interesse por essa área? 4. Você poderia elencar alguns sistemas que usuários finais utilizam no dia a dia e que possuem algoritmos inteligentes como *background?* 5. Já aconteceu alguma vez de utilizar algum desses sistemas mais populares e não entender como ele gerou determinada saída?
(iii)	**Corpo Principal de Perguntas** 6. Já aconteceu algum caso de você desenvolver um modelo e não entender a saída gerada por ele? 7. Como saber o grau de precisão de um modelo em uma sistema de *Machine Learning*? 8. A interpretação das saídas geradas por esses modelos é uma atividade fácil? 9. As saídas geradas por esses modelos são confiáveis? 10. Como identificar que esse modelo não seguiu nenhum viés? 11. Usuários finais que utilizam sistemas de ML conseguem entender como ele realiza determinadas predições? 12. Em 2016 foi aprovada pela União Européia a Regulamentação de Proteção de Dados, em que os usuários agora têm direito de explicação para ações tomadas por sistemas automatizado. Quando são desenvolvidos esses modelos, existe uma preocupação com a explicação para esses usuários? 13. Você considera que uma pessoa que não conhece nada de *Machine Learning* consegue ter um entendimento de foram gerada as predições por esses sistemas?

Fonte: elaborada pelo autor.

Aplicamos um teste piloto para que pudéssemos verificar a qualidade das perguntas, o tempo de aplicação do questionário, a identificação de viés, a ambiguidade, e se as perguntas

estavam compreensíveis para os entrevistados. Após o teste piloto, foram feitos os ajustes necessários no roteiro, como adequação de termos utilizados na entrevista.

Para realizar a análise qualitativa dos dados das entrevistas de forma exploratória utilizamos a Teoria Fundamentada nos Dados (*Grounded Theory*) (LAZAR *et al.*, 2017). A *Grounded Theory* é uma teoria indutiva baseada na análise sistemática dos dados. A ideia é que as proposições teóricas surgem dos dados obtidos na pesquisa, mais do que dos estudos anteriores. Em outras palavras, a teoria é aquilo com que o pesquisador encerra seu trabalho e não com o que principia. Não é aquilo que vai ser testado, mas aquilo que se conclui depois de uma pesquisa e da análise dos dados dela resultantes. Para dar suporte a esta etapa, utilizamos o MAXQDA[1], um *software* para análise de dados qualitativos e métodos mistos em pesquisas acadêmicas. A análise das entrevistas foi realizada utilizando uma metodologia interparticipante e intraparticipante (BARBOSA; SILVA, 2010). Nessa abordagem interparticipante, para cada pergunta foram observadas as respostas de cada participante verificando tendências centrais nas respostas. Na abordagem intraparticipante foi ouvida toda a entrevista de cada entrevistado, observando aspectos gerais e informações que foram pontuadas e que não estavam descritas nas perguntas, mas que possuem certa relevância para a pesquisa.

Os resultados desta etapa estão relatados no Capítulo 5, na Seção 5.1, onde apresentamos o processo de construção do modelo.

4.2 Uso da EngSem no contexto de explicações em IA

Nesta etapa exploramos como a EngSem, teoria que define a interação com os sistemas como um processo comunicativo entre *designers* e usuários (PRATES; BARBOSA, 2007), pode contribuir com o processo de explicações a usuários finais. Nesta teoria, a comunicabilidade é a capacidade de um *designer* conseguir transmitir para os usuários, através da interface, o *design* tal como desenvolvido por ele, incluindo o *"design rationale"* (SOUZA *et al.*, 2005).

Estudamos a teoria e alguns trabalhos fundamentados nela. Em seguida, avaliamos as interfaces de explicações sobre recomendações de publicidade do Facebook. Entendemos que as explicações da publicidade são mensagens que vieram de um processo comunicativo dos *designers* para os usuários do Facebook. Para isso, utilizamos o Método de Inspeção Semiótica (MIS)(SOUZA *et al.*, 2006), um método de avaliação por inspeção que tem por objetivo principal avaliar a qualidade da emissão da mensagem de metacomunicação do *designer* para o usuário. A

[1] https://www.maxqda.com/pt

avaliação é feita a partir da análise dos signos metalinguísticos, estáticos e dinâmicos e como eles estão sendo utilizados na interface para comunicar a mensagem do *designer*. Como o foco da inspeção está na emissão da mensagem, este método é realizado por especialistas, que percorrem a interface a fim de antecipar possíveis rupturas de comunicação que poderiam surgir na interação do usuário com o sistema.

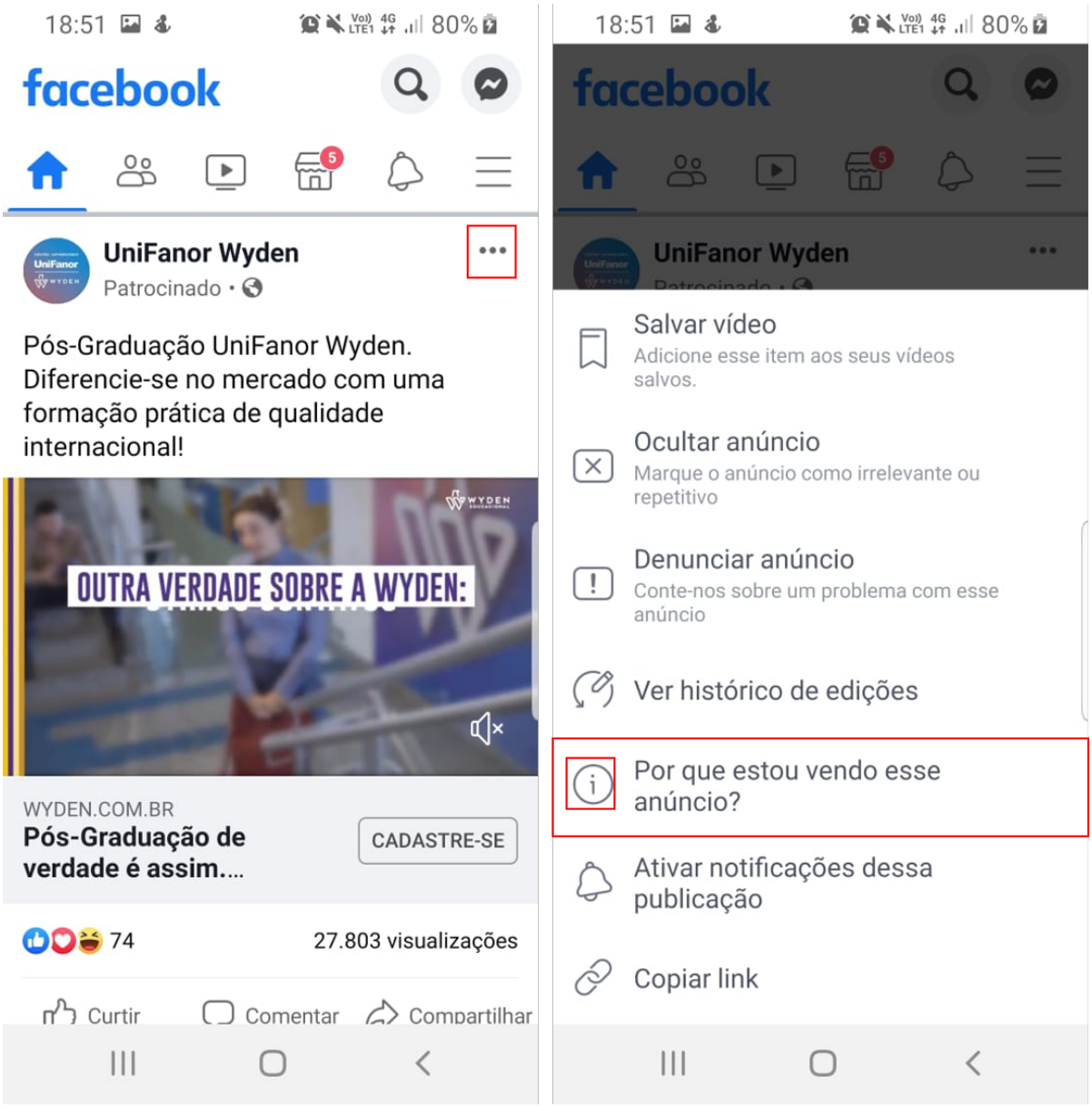
Na preparação do MIS, utilizamos como perfil de usuário uma pessoa com idade superior a 18 anos, que utiliza o Facebook regularmente através de seu *smartphone* ou computador, e que tem curiosidade em saber por que determinadas propagandas aparecem em seu *feed*. A porção do sistema analisada é composta pelas explicações contidas no menu *"por que estou vendo esse anúncio."*, obtida a partir de publicidades apresentadas na *timeline* do Facebook. Como a análise se deu especificamente na explicação fornecida, retiramos do escopo da avaliação as telas contidas no *feed* dos usuários, que correspondem ao caminho interativo para chegar a esta explicação (Figura 7).

A avaliação seguiu o seguinte cenário: *"Marcos/Maria é um(a) usuário(a) assíduo(a) do Facebook através de seu smartphone, e percebe que lhe são oferecidos alguns produtos e serviços em sua timeline. Em alguns momentos, ele(ela) não entende porque essas propagandas estão sendo exibidas. Ele(ela) então tenta encontrar explicações no Facebook, sobre porque essas propagandas estão sendo exibidas acessando o menu 'por que estou vendo esse anúncio'."*

Como complemento ao MIS, aplicamos um questionário com o objetivo de coletar as percepções dos usuários sobre as explicações. Embora a literatura (BARBOSA; SILVA, 2010) sugira a aplicação do Método de Avaliação de Comunicabilidade como complemento ao MIS, optamos pela não aplicação deste método, pois as interfaces de explicação eram compostas em grande parte por textos, dos quais o usuário realizaria somente a leitura, possuindo pouca interação.

O questionário foi elaborado no Google Formulários, tendo em vista a possibilidade de atingir um número maior de pessoas com perfis variados e com maior abrangência geográfica. O formulário continha perguntas fechadas para coleta de dados quantitativos e perguntas abertas para coleta de dados qualitativos, permitindo uma posterior análise em profundidade. O questionário foi dividido em três partes, a primeira contendo o termo de consentimento, uma segunda parte com uma coleta de dados demográficos (sexo, idade, grau de instrução) e algumas perguntas sobre o uso do Facebook, como: (1) Com que frequência você acessa o Facebook na semana? (2) Em qual plataforma acessa mais vezes (Computador/*Smartphone*/*Tablet*)? (3) Qual

Figura 7 – Captura de tela com caminho para acesso ao menu *"por que estou vendo este anúncio"* na interface do aplicativo do Facebook para acesso às explicações de recomendações de publicidade.



Fonte: Captura de tela do Facebook.

grau de relação as publicidades que aparecem na *timeline* têm com o perfil do usuário? (4) Qual o grau de interesse em entender os motivos pelo qual determinadas publicidades aparecem? (5) Alguma vez já consultou o menu “Por que estou vendo esse anúncio” no Facebook (Sim/Não)? Na terceira parte foi solicitado aos usuários que realizassem uma pequena tarefa acessando o conteúdo do menu “Por que estou vendo esse anúncio”. Nesta tarefa, eles deveriam explorar o conteúdo da explicação e tentar entender por que o Facebook apresentava aquela publicidade. Após essa tarefa deveriam responder a algumas perguntas, como: (6) Que plataforma usou para realizar a tarefa (Smartphone/Computador)? (7) Qual o grau de compreensão da explicação

apresentada? (8) A explicação era suficiente para entender os motivos do anúncio ser apresentado no *feed*? (9) Qual o grau de entendimento dos ícones apresentados na explicação? (10) Qual o grau de entendimento das palavras em negrito apresentados na explicação? (11) Qual nota o usuário dá para a explicação apresentada? Além dessas perguntas, o usuário poderia enviar uma captura das telas nas quais ele leu a explicação e responder a uma pergunta aberta se ainda tinha alguma dúvida que gostaria de ser esclarecido sobre ela. Algumas perguntas (3, 4, 7, 8, 9 e 10) eram respondidas seguindo uma escala de 1 a 8. As perguntas da segunda parte continham campos para justificação da resposta de forma opcional para os usuários.

Foi realizado um teste piloto para validar a qualidade das perguntas, o entendimento dos usuários e o tempo de resposta. O teste foi realizado com uma pessoa da área de tecnologia e outra que não era da área. O tempo de resposta do teste variou entre 8 e 15 minutos. Foi perguntado aos usuários se as perguntas estavam compreensíveis e se os termos utilizados como *"printar"*, *"feed"* ou *"timeline"* eram compreensíveis. Ao final do teste piloto, foi observado que a pergunta relacionada à frequência de uso do Facebook deveria ser alterada para semanal, ao invés de diária.

Para realizar a análise qualitativa dos dados do formulário, nos baseamos na Teoria Fundamentada nos Dados (*Grounded Theory*) (LAZAR *et al.*, 2017) com o auxílio da ferramenta MaxQDA[2] na versão 2018.

O relatório final deste experimento pode ser encontrado no seguinte endereço eletronico: encurtador.com.br/luL46. Este trabalho foi publicado no Anais XVIII Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais em 2019 (CARVALHO *et al.*, 2020).

Além da avaliação, após os resultados obtidos, realizamos uma revisão de literatura com o objetivo de aprofundar os conhecimentos sobre os modelos, os métodos de avaliação e as ferramentas epistêmicas propostas na literatura, fundamentadas na Engenharia Semiótica, e analisar o quanto esses recursos podem apoiar o problema de explicações a usuários finais no contexto de IA. Durante essa revisão foram levantados quais desses recursos promovem a reflexão sobre aspectos sociais, tendo em vista que os usuários finais de sistemas inteligentes, que utilizam dados de uso para gerar predições, têm grandes preocupações com questões éticas. Os estudos realizados nesta etapa foram apresentados de forma resumida na Seção 2.2.

[2] https://www.maxqda.com/pt

4.3 Definição do MoReXAI

Nesta etapa, formalizamos o MoReXAI, primeiramente identificando os tópicos de discussão que seriam utilizados no modelo e definindo cada elemento e as etapas para sua aplicação.

As explicações dos modelos de aprendizagem de máquina a usuários finais estão diretamente ligada aos objetivos dos usuários. Na Seção 2.1.2 observamos que boa partes desses objetivos estão relacionados com princípios éticos e busca por desenvolvimento de IA mais responsável. Esta etapa da metodologia consistiu em realizar revisão bibliográfica em diretrizes de princípios éticos em órgãos governamentais, organizações da sociedade civil e iniciativa privada na busca pelos tópicos de discussão que seriam acrescentados ao MoReXAI. Na Seção 5.3.1 especificamos detalhadamente como foram selecionados os tópicos e as perguntas que serão utilizadas no modelo.

Após esta etapa formalizamos o modelo definindo seus elementos. O MoReXAI é baseado no modelo de comunicação de Jakobson (1960) apud Santee e Temer (2011), possuindo os mesmos elementos: emissores, receptores, canal, mensagem, contexto e código. Aqui também formalizamos as perguntas, associado-as a cada tópico, que devem guiar a discussão. Além disso, descrevemos as etapas de uso do modelo: Planejamento, aplicação do modelo e análise de resultados, os quais serão detalhados na Seção 5.3.

4.4 Estudo de caso

Realizamos um estudo de caso exploratório para observar o uso do modelo em um contexto real de desenvolvimento de um sistema de IA.

O estudo foi realizado durante o desenvolvimento de um sistema de recomendação de serviços do Governo do Estado do Ceará que funcionará dentro do aplicativo Ceará App. O Ceará App é um aplicativo *mobile* com objetivo de ofertar os principais serviços do governo de forma rápida e concentrada em um único local de forma remota. Através dele os usuários podem acessar serviços como: plantão *on-line* para atendimento de profissionais de saúde 24h, agendamento de teste Covid-19 e vacinação, emissão de certidões negativas e de regularidade, requerimento da carteira de habilitação (CNH) e solicitação de 2ª via ou renovação da CNH, anúncio, busca e compra produtos da agricultura familiar, entre outros.

5 O MOREXAI

A definição do MoReXAI se deu forma exploratória e, cada etapa realizada da metodologia, foi agregando conhecimento e *insights* para a construção do MoReXAI. Neste capítulo apresentamos o processo de construção do modelo, no qual inicialmente obtemos a visão dos especialistas sobre explicações a usuários finais no contexto de XAI (Seção 5.1). Depois exploramos como a Engenharia Semiótica pode contribuir com o processo de explicações a usuários finais, e, além das revisões de literatura, realizamos uma avaliação de explicações sobre recomendações de publicidade do Facebook utilizando o MIS (Seção 5.2). Depois iniciamos o processo de formalização do MoReXAI, inicialmente buscando tópicos de conversas e perguntas que fariam parte do modelo, depois definindo os elementos e o fluxo de aplicação do MoReXAI (Seção 5.3).

5.1 Obtendo visão dos especialistas sobre explicações em IA

Iniciamos o processo de construção do modelo realizando entrevistas com especialistas em IA com o objetivo de entender como eles pensam explicações a usuários no contexto de desenvolvimento de modelos de aprendizagem de máquina.

Foram realizadas um total de seis entrevistas. Todos os entrevistados possuíam experiência no desenvolvimento de sistemas de aprendizado de máquina, e na realização e orientação de trabalhos acadêmicos na graduação, mestrado e doutorado. O tempo de experiência com aprendizado de máquina, variava de 6 meses a 12 anos, nas áreas de visão computacional, aprendizado por reforço, sistemas de recomendação, clusterização e testes em ML. Os entrevistados possuíam trabalhos com desenvolvimento de modelos de ML em diversas áreas como: segurança pública, tráfego urbano, agronomia, educação e jogos. A Tabela 3 apresenta o perfil dos entrevistados de forma sintetizada.

5.1.1 Impacto na vida das pessoas

De acordo com os entrevistados, os sistemas de ML rodam como *background* em várias aplicações no dia a dia de forma transparente para usuários. Foram citados sistemas como: busca do Google, *marketing* digital e reconhecimento facial em redes sociais como Facebook e Instagram, aplicações de reconhecimento de voz e teclado virtual em celulares Android e Apple, transcrição de textos automático no Youtube, recomendação de produtos em sistemas de

Tabela 3 – Perfil dos entrevistados com tempo de experiência e áreas em que desenvolvem pesquisas relacionadas a *Machine Learning*

Entrevistado	**Tempo**	**Área**	**Experiência em ML**
E-01	7 anos	Visão Computacional e Clusterização	Agricultura, Financeiro, Segurança Pública e Tráfego Urbano
E-02	2 anos	Sistemas de Recomendação	-
E-03	6 meses	Testes em ML	-
E-04	3 anos	Visão Computacional	Tráfego Urbano, Segurança Pública, Agricultura, Financeiro
E-05	3 anos	Visão Computacional e Aprendizado por Reforço	Sistemas Multiagentes, Educação e Financeiro
E-06	12 anos	Visão Computacional	Educação e Financeiro

Fonte: elaborada pelo autor.

comércio eletrônico, reconhecimento de *spam* em *e-mails*, entre outros. Por estarem inseridos no cotidiano das pessoas de forma pervasiva, é consenso entre eles que esses sistemas afetam a vida dos usuários. Esses impactos têm características positivas quando, por exemplo, podem usar essa tecnologia para melhorar a vida das pessoas em áreas como agricultura, medicina, segurança pública, comércio, ou seja, quando automatizam processos, e realizam funções que algumas vezes são difíceis para os seres humanos. Observamos esse pontos em depoimentos como: **E-06:** *"(...) hoje algoritmos que fazem análises de imagens médicas (...) já tem casos de algoritmos que fazem diagnostico melhor (...) com mais precisão que seres humanos."*; e **E-04:** *"(...) de algum modo é para melhorar a qualidade de vida, o uso do tempo, uso dos recursos, ou gastar menos combustível ou gastar menos energia, gastar menos produtos nos processos competitivos, onde as empresas estão reduzindo custos, então a intenção é sempre diminuir custos, otimizar processos, detectar fraudes (...)".* Alguns desenvolvedores também reconhecem que pode haver, em alguns momentos, impactos negativos desses sistemas, principalmente quando os modelos utilizados foram construídos a partir de uma base de dados pequena e não tão abrangente, podendo causar predições erradas ou enviesadas. Neste contexto, citaram como exemplos, incentivo a compra de produtos e criação de "bolhas", em que, por exemplo, os algoritmos passam a decidir quais filmes ou vídeos os usuários vão assistir, baseado nas suas preferências e seu histórico, não apresentando assim outros conteúdos diversificados, como podemos ver no depoimentos dos entrevistados: **E-03:** *"(...) meio que tentam induzir você a fazer algo, se você está pesquisando aquilo, mas ela indica que isso aqui está barato, tendencia você a comprar, ou oferecer um serviço da mesma forma. (...)"*; e **E-02:** *"(...) a pessoa começa a perder a noção de verdade, pois dá a um público um conteúdo que não necessariamente o conteúdo é relevante se você parar pra pensar, mas por acaso o algoritmo chegou à conclusão que era. E também um exemplo clássico*

na criação de bolhas.".

Os entrevistados reconhecem que o uso desses sistemas, em algumas situações, pode apresentar predições erradas, como: propagandas que estão fora do contexto em redes sociais e aplicações de comércio eletrônico, *e-mails* enviados para a caixa de *spam* de forma errada, recomendações de filmes que não se encaixam no perfil do usuário, reconhecimento de imagens que não condizem com a apresentada. Quando perguntados a respeito da confiança em sistemas que usam ML, os entrevistados responderam que variam de acordo com o quão sensível é esse sistemas. Aplicações que têm interferência direta na vida das pessoas, ou que envolvem negócios e riscos financeiros por parte de empresas, tendem a necessitar de um grau de precisão maior, para que melhore a confiança por parte dos usuários. Em contrapartida sistemas que não são sensíveis e que funcionam de forma transparente, geralmente, não despertam interesse dos usuários em saber como estes sistemas funcionam, fazendo com que o uso deles seja feito, sem se preocupar com a precisão das saídas. Segundo os especialistas, as informações a respeito de como os algoritmos funcionam despertam interesse somente de pessoas da área de tecnologia. Podemos perceber este pensamento em depoimentos como: **E-02:** *"(...) pode-se ter aquele consumidor mais exigente que vai querer entender como funciona, mas muitos compram por conta da feature mesmo, porque fazem aquela predição. Eu vou comprar o iphone, porque tem a siri que fala. Tem gente que prefere depois de um certo uso, a assistente do google ou a assistente da apple, porque pra determinada situação funciona pra outras situações não funciona."*; ou **E-04:** *"Por exemplo o reconhecimento de imagem você bota o celular na sua frente e ele automaticamente entra, reconhece sua face, e esse reconhecimento se dá por essas técnicas também, e o usuário acha aquilo a coisa mais natural do mundo, não está nem ai pra saber o que há por traz disso"*; e **E-01:** *"(...) se você for negado um empréstimo você quer saber por que foi negado o empréstimo, agora se ele fez uma recomendação errada no filme, você pode não se importar ou você pode querer saber por que, depende do contexto né (...)"*.

5.1.2 Interpretabilidade dos modelos de ML

Quanto à interpretabilidade dos modelos, os entrevistados citaram que, no processo de desenvolvimento, alguns sistemas são mais fáceis de interpretar (modelos que utilizam algoritmos de classificação e árvores de decisão), por se tratarem de algoritmos paramétricos, ou que possuem funções de custo que facilitam o processo de interpretação, como em: **E-02:***"Dependendo do modelo a gente pode avaliar a função de custo, (...) se a gente tem alguma*

função que ajude a avaliar o modelo e é um modelo paramétrico a gente pode ir mudando os parâmetros e ver se ele vai melhorando ou piorando ao longo do tempo (...). Mas não é todo modelo que permite, tem uns que são bem complicados de entender, tem uns que tem são muito custosos, tipo modelos de deep learning (...).". No entanto, alguns modelos apresentam um maior grau de complexidade (*deep learning*) no processo de interpretação das saídas, sendo necessário para isso definir métricas que auxiliem na interpretabilidade, ou contar com a ajuda de especialistas de domínio, como observado em: **E-06:** *"(...) por exemplo reconhecimento de imagem, você ter que classificar uma imagem (...) não é uma tarefa fácil a interpretabilidade, por que o modelo classificou daquela forma nem sempre isso fica claro, pode ser uma caixa preta, o modelo pode fazer bem, só que não fica muito fácil de entender como é que o modelo conseguiu chegar naquele resultado. (...) o aprendizado é automático, você define uma serie de camadas de aprendizado, o pessoal chama isso de aprendizado profundo, deep learning (...)"*; ou em **E-01:** *"(...) não existe uma verdade absoluta sobre como criar clusters, existem métricas que provam que o cluster pode ser (...) pouco compacto, que significa o que né, os objetos não tem tanta similaridades assim como você está dizendo, mas não há uma verdade absoluta, é muito subjetivo, então existem problemas que têm esse embasamento verdade absoluta e vão existir problemas que não têm. E esses problemas que não têm ai você tem que trabalhar com as métricas ou com especialistas de domínio, aí é mais difícil."*. Para os desenvolvedores, esta tarefa de interpretar modelos de ML é desafiadora, e não é uma atividade fácil, mas a experiência no desenvolvimento desses sistemas ajuda a fazer os ajustes necessários para sua otimização, principalmente quando se tratam de modelos que utilizam algoritmos caixas-pretas, como as redes neurais, como observado em: **E-04:** *"Existem momentos em que você fica se perguntando por que, e fica tentando usar o que a gente chama um pouco de intuição. (...) Claro quando você vai adquirindo muita experiência em uma área você vai tendo uma intuição de por que aquele algoritmo deu aquela saída né."*.

De acordo com os entrevistados, para termos algoritmos mais precisos e com maior acurácia, precisamos de um volume de dados cada vez maior. Vale ressaltar que alguns entrevistados citaram que a expertise do especialista no desenvolvimento desses sistemas ajuda na identificação de viés, assim como no refinamento das entradas, para que os modelos melhorem seus resultados. Apenas um dos desenvolvedores considerou a participação de usuários finais no processo de desenvolvimento dos modelos, como uma forma de melhorá-los. **E-06:** *"(...) acho que ter a possibilidade de saber a explicação sempre é bom né, se for possível entender por*

que ele chegou naquela conclusão, parece ser uma coisa boa, alguns usuários dependendo do contexto vão se interessar e outros não." ou **E-01:** *"(...) eu acho que se as pessoas soubessem, elas poderiam até ajudar mais, (...) ter um human-in-the-loop. Se eles tiverem explicação talvez eles queiram ajudar mais."*

5.1.3 Explicações dos modelos de ML

Os desenvolvedores se deparam com a dificuldade de explicar esses sistemas que utilizam *Machine Learning*, embora considerem importante e bastante desafiador explicar o funcionamento desses modelos para usuários, que muitas vezes não têm conhecimento da área. Consideram que as explicações para usuários e especialistas de domínio, está no âmbito somente de explicações locais, não tendo muito interesse em entender como é o funcionamento do algoritmo ou toda a fundamentação matemática existente para a realização da predição gerada através de explicações globais. Em contrapartida, percebe-se também que, para os especialistas, o GDPR não vai ser totalmente implementado, tendo em vista que as empresas não têm interesse em revelar suas estratégias de negócio, como observado por (WACHTER *et al.*, 2017), o que pode ser corroborado pelo pensamento dos entrevistados em **E-05:** *"(...)a empresa não pode dizer por que está revelando uma estratégia de negócio. Utiliza tais algoritmos para ter maiores detalhes, a empresa não vai se dar o trabalho de explicar como funcionar seus algoritmos(...)"*. No entanto, foi citado que a regularização do direito a explicação pode afetar, de forma bastante rigorosa, na escolha dos algoritmos para o desenvolvimento em sistemas de *machine learning*, tendo em vista a dificuldade de explicar modelos caixas-pretas, como observado em: **E-06:** *"(...) então isso talvez impacte em alguns modelos mais difíceis de extrair explicação, talvez eles passem a ser menos utilizados ou não utilizados pela dificuldade de extrair explicação dali. (...) geralmente o objetivo é maior acurácia, (...) o fato de ser facilmente explicável não acredito que seja um critério comum nesses projetos de aprendizado de maquina, mas passando a ser obrigação isso vai afetar ate mesmo a escolha dos modelos, alguns modelos podem cair em desuso pela dificuldade da explicação."*

Perguntamos aos entrevistados se, no processo de desenvolvimento de sistemas de ML eles, em algum momento, pensavam em explicações a usuários. A maioria, respondeu que geralmente não se pensa em explicar aos usuários a respeito funcionamento dos algoritmos, ou como eles chegaram a determinada predições, e que o principal interesse estava na melhoria das métricas utilizadas nos modelos. Para os desenvolvedores, quando se pensa em melhorar a

qualidade dos modelos de ML, automaticamente se está pensando nos usuários que vão utilizar esses sistemas, e que, quanto mais essas tecnologias avançam e se tornam mais pervasivas coletando mais dados dos usuários, maior a precisão e acurácia de seus resultados.

A partir dos estudos realizados nesta etapa, percebemos que, sob a ótica dos especialistas em ML, existe uma certa dificuldades na interpretabilidade dos modelos que utilizam aprendizado de máquina, principalmente quando se tratam de algoritmos caixas-pretas. Assim, se faz necessário para uma melhoria desse processo, um aumento considerável na quantidade de dados utilizados para treinar os modelos e refinamentos dos parâmetros, bem como para adotar métricas, e incluir especialistas de domínio, para auxilio na interpretação dos modelos gerados. Nesse sentido, notamos que a principal preocupação dos desenvolvedores está em melhorar a acurácia dos modelos, durante a etapa de desenvolvimento. Segundo os especialistas, o principal beneficiário do aumento da acurácia dos modelos são os próprios usuários que utilizam esses sistemas. Para os desenvolvedores, não existe interesse por parte dos usuários, entender como esses algoritmos realizam determinadas predições, principalmente quando se tratam de sistemas mais simples, e que não são sensíveis, como filtro de *spam*, corretor ortográfico, entre outros. No entanto, quando esses sistemas têm impactos diretos e que incorrem em riscos mais graves, existe uma tendência dos usuários buscarem explicações a respeito dos motivos de determinadas predições. Explicar esses algoritmos, segundo eles, é algo bastante complexo devido à grande quantidade de conceitos matemáticos e computacionais que fogem a compreensão dos usuários. Embora não exista uma preocupação direta com usuários, por parte dos desenvolvedores, no momento do desenvolvimento, eles consideram importante ter usuários ajudando, mesmo que no processo de avaliação dos modelos. Ter um *human-in-the-loop* corrigindo imprecisões nas previsões de máquinas auxilia no aumento da precisão, o que resulta em maior qualidade dos resultados (MUNRO, 2019). Além disso é importante também pensar estratégias para incluir usuários tanto no processo de interpretabilidade quanto na construção das explicações desses modelos.

Este etapa da metodologia não buscou apresentar maneiras de explicar sistemas de IA, mas explorar como especialistas pensam explicações para usuários. Para os desenvolvedores a dificuldade de entendimento dos conceitos matemáticos relacionados aos algoritmos de aprendizagem de máquina podem afetar o entendimento dos usuários quanto às explicações e consideram a participação deles somente validando as predições geradas pelos modelos. No entanto, uma explicação não precisa traduzir toda a lógica envolvida no seu processo de infe-

rência, ela deve responder a questões próprias dos usuários, de forma que melhore a confiança nesses sistemas (EIBAND *et al.*, 2018b). Assim, como observado por Carbonera *et al.* (2018), fazer uso da Engenharia Semiótica para endereçar o problema de explicações, como um processo comunicativo, e que o designer e usuários estão integrados nesse processo através das teorias de construção de IA e teorias de uso, podem ser um caminho para explicações mais efetivas, satisfazendo às perspectivas dos usuários e legislações vigentes.

Em pesquisa similar realizada por Brennen (2020), os autores perceberam que há uma inconsistência na terminologia para discutir IA explicável por parte dos *stakeholders*. A falta de consistência na terminologia atrapalha a discussão, pois permite que as pessoas falem de algo sem compreenderem, muitas vezes, o que estão fazendo. A comunicação é agravada pelo uso de palavras que têm uma definição (ou múltiplas definições) e o uso coloquial totalmente diferente. Observaram que a IA explicável incorpora vários objetivos distintos, que são importantes para pessoas diferentes por motivos diferentes. Por exemplo, as motivações para obter explicação de como o algorítimo de IA funciona podem estar relacionadas à depuração dos modelos, à detecção de viés e à construção de confiança nos modelos. Concluem que a falta de terminologia consistente relacionada aos conceitos e objetivos da IA explicável atrapalham a discussão, especialmente entre pessoas com diferentes origens disciplinares. Enfatizam que, para progredir nesta área, é necessária a definição clara de termos-chave, para ajudar as partes interessadas a se comunicarem de maneira mais eficaz (BRENNEN, 2020).

5.2 Avaliação das explicações em recomendações de publicidade do Facebook sob a ótica da EngSem

Na segunda etapa da metodologia, investigamos como a EngSem pode contribuir com o processo de explicações em IA, para isso realizamos um estudo exploratório com o objetivo de avaliar as explicações sobre recomendações de publicidade do Facebook, utilizando o Método de Inspeção Semiótica e Questionários. Os resultados desta pesquisa foram publicados no XVIII Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (CARVALHO *et al.*, 2020).

5.2.1 Resultado de aplicação do MIS

Realizamos a análise segmentada dos signos metalinguísticos, estáticos e dinâmicos e ao final reconstruímos a metamensagem do *designer* para cada signo.

5.2.1.1 Análise segmentada dos signos metalinguísticos

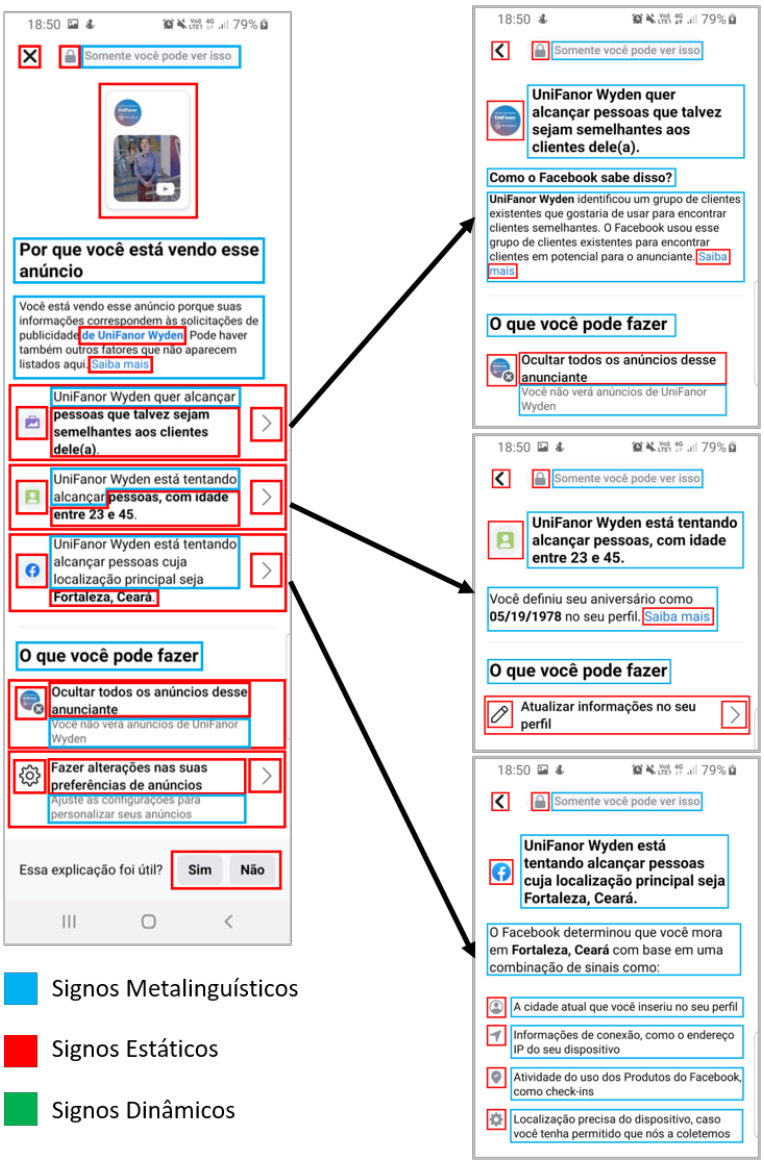
Quem é o usuário visado pelo *designer*? O que sabe e conhece (ou não)? Os usuários que buscam essas explicações são aqueles que estão interessados em saber por que estão vendo determinada publicidade (*"Por que você está vendo este anúncio"*) e que são público-alvo das empresas anunciantes (*"Você está vendo este anúncio porque suas informações correspondem às solicitações de publicidade de..."*), como podemos ver nos signos metalinguísticos da Figura 8. Os usuários estão interessados em explicações simples, ou explicações mais detalhadas e, como eles podem não conhecer alguns ícones apresentados na interface, são adicionados signos metalinguísticos para explicar esses ícones. **O que ele quer ou precisa fazer, de que formas preferenciais (potencialmente alternativas), e por quê?** A mensagem do *designer* para o usuário sobre o que ele pode fazer está bem explícita no signo metalinguístico *"O que você pode fazer"* (Figura 8). O usuário pode ocultar o anúncio ou ajustar as configurações para personalizar as publicidades apresentadas. Para obter explicações detalhadas, o usuário pode acessar os tópicos da interface de explicação simplificada que conterá os motivos da exibição do anúncio em sua *timeline* (Figura 8).Se necessitar conhecer os fatores e regras que influenciam a exibição do anúncio, pode acessar o *link "Saiba mais"*. Esta explicação contém termos em azul com explicações mais específicas relacionadas a esses termos (Figura 9). **O que é o sistema produzido com este usuário em mente e como pode ou deve ser usado para realizar um conjunto de objetivos ou efeitos contemplados pelo *designer*?** O conteúdo do *"por que estou vendo este anúncio"* é um grande signo metalinguístico. Através de textos, ele explica os motivos pelos quais os usuários estão vendo um determinado anúncio. A interface foi pensada em apresentar essa explicação de forma gradual a partir de navegação hierárquica, permitindo que os usuários possam se aprofundar nesta explicação, a partir de seu interesse. Assim, inicialmente é apresentado pelo *designer* uma interface simplificada, contendo algumas informações que influenciam na exibição do anúncio como idade, localização e semelhança do perfil do usuário com clientes do anunciante e que ele está vendo determinado anúncio porque o perfil dele tem forte relação com o perfil procurado pelo anunciante. O usuário pode

obter mais informações a partir do menu *"Saiba mais"*, que contém as regras utilizadas para a recomendação de publicidade, apresentando em destaque (*link* em azul) algumas palavras que podem gerar dúvidas aos usuários e que podem ser explicadas em mais detalhes a partir de sua interação com esse *link* (Figura 9). O Facebook utiliza algumas categorias que influenciam na exibição dos anúncios, como: Atividades nos Produtos Facebook (interesses, categorias, públicos semelhantes), atividades com outras empresas (públicos personalizados, atividade *off-line*), informações de localização. O *design* pensou em uma tela intuitiva etiquetando a maioria dos ícones apresentados na interface com signos metalinguísticos (Figura 7). Como muitos dos dados são privados, ele torna a tela da explicação privativa somente ao usuário-alvo do anúncio (Figura 8).

5.2.1.2 Análise segmentada dos signos estáticos

Quem é o usuário visado pelo *designer*? O que sabe e conhece (ou não)? O usuário é alguém que conhece o aplicativo do Facebook ou tem facilidade em navegação hierárquica na *web* e em *smartphone* e, a partir de *links* e ícones, pode avançar e voltar nas interfaces de explicação. **O que ele quer ou precisa fazer, de que formas preferenciais (potencialmente alternativas), e por quê?** Para fazer a leitura da explicação, o usuário precisa navegar pela tela e, sempre que quiser mais detalhes, deve interagir com as informações apresentadas. O detalhamento pode ser conseguido a partir do item "Saiba mais" ou os tópicos de explicação que possuem um botão para avançar e obter um detalhamento deles. Boa parte dos signos estáticos possui um signo metalinguístico associado, como no signo "Ocultar todos os anúncios desse anunciante" (Figura 8), em que ele explica o que vai acontecer caso o usuário resolva realizar essa ação. O usuário também pode informar se a explicação foi útil ou não. O usuário pode interagir com determinadas partes do texto que se apresentam em destaque (azul), para obter mais detalhe a respeito daquela palavra ou daquele trecho de palavras (Figura 9). **O que é o sistema produzido com este usuário em mente e como pode ou deve ser usado para realizar um conjunto de objetivos ou efeitos contemplados pelo *designer*?** A explicação produzida exibe inicialmente uma imagem relacionada a empresa anunciante e em seguida alguns signos estáticos que estão associados a categorias usadas para relacionar aquele anúncio ao perfil do usuário. A comunicação através da interface foi construída de forma que o usuário pode ir adquirindo explicações mais detalhadas a partir da navegação pelos signos estáticos. Assim, o usuário pode acessar esse detalhamento a partir da palavra "Saiba mais", ou acessando os tópicos

Figura 8 – Exemplos de signos metalinguísticos e estáticos exibidos nas explicações de recomendações de publicidades do Facebook.



Fonte: Captura de tela de explicações sobre recomendações de publicidade do Facebook.

de informações exibidas nas quais se basearam sua recomendação. Caso o usuário deseje, pode ocultar o anúncio, ou fazer alterações referentes às suas preferências de exibição de publicidade ou seu perfil. Pode também avaliar se a explicação é útil. Essas informações apresentadas são privadas e somente o usuário pode visualizá-las.

5.2.1.3 Análise segmentada dos signos dinâmicos

Quem é o usuário visado pelo *designer*? O que sabe e conhece (ou não)? O usuário deseja conhecer mais a respeito das regras utilizadas para apresentar determinadas publicidades em sua *timeline* ou conhecer com mais profundidade como o Facebook sugere as

Figura 9 – Captura de tela de parte da interface do menu "Saiba mais"com explicações das regras e dos fatores que influenciam as recomendações de publicidades do Facebook.

Por que estou vendo anúncios de determinado anunciante no Facebook?

Compartilhar artigo

Queremos que os anúncios que você vê no Facebook sejam personalizados de acordo com seus interesses e lhe apresentem novidades. Veja algumas das informações que usamos para exibir anúncios que acreditamos ser interessantes a você:

Sua atividade nos produtos do Facebook:

Algumas das suas atividades nos Produtos do Facebook podem nos ajudar a exibir anúncios que acreditamos ser interessantes a você. As informações que podemos usar para exibir anúncios a você incluem:

Interesses e categorias. Algumas informações que você nos fornece podem nos ajudar a pensar que está interessado em algo, como culinária ou boa forma, ou que faz parte de um grupo maior (chamado de categoria), como um usuário móvel. Podemos usar seus interesses e suas categorias para conectá-lo a empresas que acreditamos ser interessantes a você.

Informações que podem nos ajudar a adicionar você a um interesse ou a uma categoria:

- Páginas do Facebook que você curtiu.
- Informações do seu perfil do Facebook e do Instagram.
- Anúncio em que você clicou.
- Locais em que você fez check-in usando o Facebook.

Públicos Semelhantes. O Público Semelhante é um dos recursos do Facebook que os anunciantes usam para alcançar pessoas que possam ser similares às pessoas que já mostraram interesse nas empresas deles. Usamos algumas das ações que você realiza ou informações que nos fornece para ajudar a criar os Públicos Semelhantes.

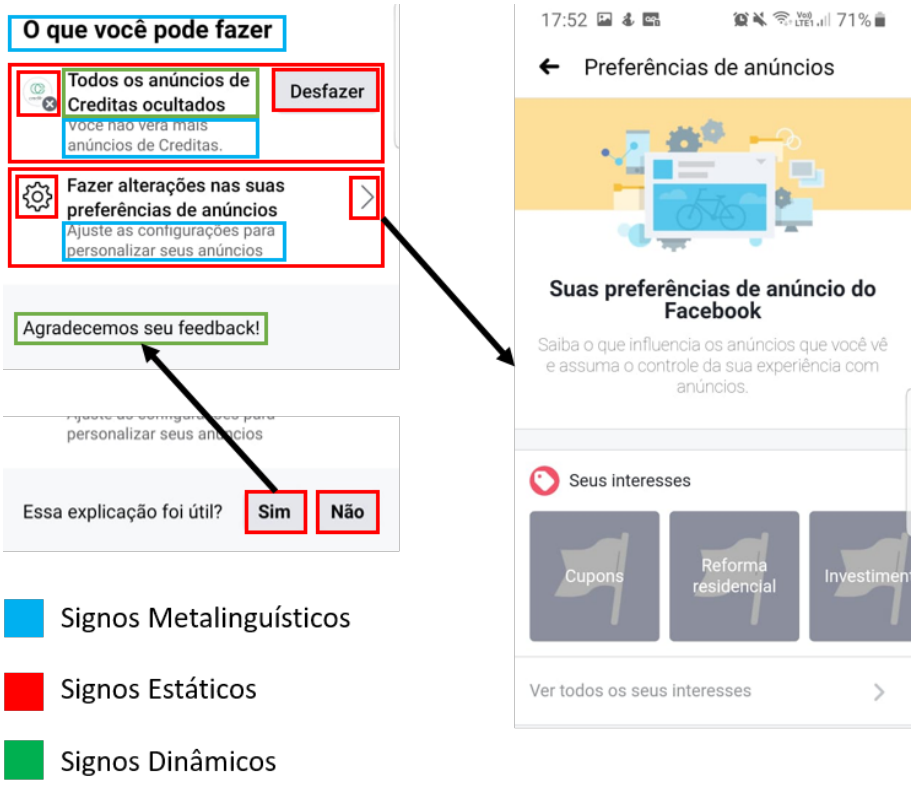
Por exemplo, os anunciantes podem nos dizer que desejam exibir anúncios a pessoas que possam ser semelhantes àquelas que já se interessaram na empresa deles (por exemplo, pessoas que curtiram a Página ou acessaram o site deles). Fazemos a correspondência de informações como interesses ou dados demográficos (por exemplo, sua cidade) das pessoas que já se interessaram na empresa para ajudar a exibir anúncios a outras pessoas que possam estar interessadas na empresa ou nos produtos do anunciante.

Fonte: Captura de tela de explicações detalhadas sobre recomendações de publicidade do Facebook.

publicidades (Figura 8). Deseja também dar *feedback* se a explicação foi útil, ou ocultar anúncios de determinado anunciante. **O que ele quer ou precisa fazer, de que formas preferenciais (potencialmente alternativas) e por quê?** Conhecer com mais detalhes as regras aplicadas para personalizar publicidades no Facebook. Lendo as informações contidas na explicação e navegando em textos em formato de *link* para entender mais a respeito deste conteúdo. Os usuários podem também ocultar anúncios do anunciante e avaliar a explicação dada a ele do por que de estar vendo determinado anúncio (Figura 10). O usuário pode fazer alterações em suas preferências de anúncio. E avaliar se uma determinada explicação foi útil. **O que é o sistema produzido com este usuário em mente e como pode ou deve ser usado para realizar um conjunto de objetivos ou efeitos contemplados pelo *designer*?** A partir da interação com alguns signos estáticos, o usuário pode detalhar as explicações. Pode também avaliar a explicação em que recebe uma mensagem de agradecimento do *designer* pela avaliação realizada.

Pode também ocultar anúncios e desfazer quando quiser, seja através da interface apresentada na Figura 10, ou a partir das suas alterações nas suas preferências de anúncios.

Figura 10 – Exemplos de causa em signos dinâmicos exibidos na interface de explicação das recomendações de publicidades do Facebook.



Fonte: Captura de tela de explicações sobre recomendações de publicidade do Facebook.

5.2.1.4 Reconstrução da metamensagem de cada signo

Reconstruímos a metamensagem de cada classe de signo conforme descrito a seguir:

– **Signos Metalinguisticos:** *"Acredito que você é um usuário que utiliza o Facebook e está querendo entender os motivos de estar vendo determinadas publicidades no Facebook. Em alguns momentos você se dará por satisfeito com uma explicação simples, em outros terá necessidade de uma explicação detalhadas das regras utilizadas para exibição desses anúncios. Portanto, esta é a explicação que projetei para você. Em uma primeira interface você visualiza algumas informações utilizadas para recomendar publicidades para você em um formato de texto simples, separada por tópicos e, caso você necessite conhecer mais sobre o assunto, reservei um espaço com explicações detalhadas de como recomendamos suas publicidades. As explicações são apresentadas em formato de texto e tento utilizar uma linguagem simples para facilitar o seu entendimento. Etiquetei quase todos os*

elementos da interface com signos estáticos para que você não sinta dificuldades em entender como ela deve ser usada. Você pode, sempre que quiser, ocultar a exibição de publicidade do anunciante, ou alterar suas preferências de anúncio para que possamos recomendar publicidades que se encaixam em seu perfil. Pode também avaliar se a explicação fornecida a você foi útil. Como as informações usadas na explicação são privadas tivemos o cuidado para que somente você tenha acesso a essas informações. No detalhamento aprofundado das explicações contidas no menu 'Saiba mais', dividi as informações utilizadas para recomendar publicidades em categorias, com a possibilidade de ainda poder verificar termos específicos relacionados a essas categorias. Tento sempre que possível incluir textos explicativos relacionados aos ícones e funcionalidades da qual você, usuário, pode estar utilizando. Caso você queira, você pode compartilhar essas informações."

– **Signos Estáticos:** *"Acredito que você é um usuário frequente do Facebook ou tem familiaridade com páginas de internet ou navegação entre telas no aplicativo do Facebook. Como você está querendo entender os motivos de estar vendo determinadas publicidades, projetei uma interface com explicações a respeito destas recomendações. Assim, em negrito você terá uma explicação simples, com algumas informações que foram usadas para recomendar uma publicidade e caso você tenha interesse em informações mais detalhadas basta navegar através do 'Saiba mais', ou clicar em uma dessas informações destacadas. Incluí alguns ícones ao lado desses tópicos que fazem referências a categorias utilizadas na hora de recomendar a publicidade. Estas categorias são melhor entendidas a partir da leitura do conteúdo do 'Saiba mais'. Algumas palavras ou trechos de palavras contidas na explicação detalhada do saiba mais estão em negrito ou em azul, para informar que você pode obter mais informações a respeito destes elementos. Grande parte das funcionalidades possuem textos explicativos para facilitar o seu entendimento. Para melhorar a sua navegação ou algumas ações que você pode realizar nas interfaces utilizei signos convencionais como: avançar, voltar, fechar, links e botões de ação, para que você não sinta dificuldade em navegar/usar a interface. Você pode ocultar as publicidades que não condizem com o seu perfil ao interagir com o signo 'Ocultar anúncios deste anunciante', ou alterar suas preferências de visualização de anuncio. Você também pode avaliar se a explicação foi útil respondendo 'sim' ou 'não' a essa pergunta. Além disso você também pode acessar suas configurações de perfil, quando o anúncio tiver*

alguma relação com esta informação, e alterar suas informações. Deixei as informações apresentadas na explicação privadas, assim somente você terá acesso a elas. Se quiser pode compartilhar a explicação aprofundada contida no 'Saiba mais'."

– **Signos dinâmicos:** *"Acredito que você é um usuário que está querendo entender os motivos de estar vendo determinadas publicidades no Facebook. Em alguns momentos você se dará por satisfeito com uma explicação simples, em outros terá necessidade de uma explicação detalhadas das regras utilizadas para exibição desses anúncios. A partir da interação com as explicações você poderá obter mais detalhes clicando nas informações apresentadas ou clicando no link 'Saiba mais'. Sempre que você resolver ocultar anúncios de determinando anunciante, o signo estático referente a esta funcionalidade será alterado informando que todos os anúncios foram ocultados e você pode desfazer essa ação através do botão 'Desfazer'. Ficarei muito agradecido sempre que obter seu feedback sobre a explicação, informando se ela é útil ao clicar em 'sim' ou 'não' à pergunta que faço na interface."*

5.2.1.5 Alinhamento e comparação

Ao integrar as três mensagens de metacomunicação segmentada, notou-se a falta de alguns conteúdos do *template* de metacomunicação (incompletude) que a Engenharia Semiótica define como modelo da mensagem do *designer* para o usuário. A identificação do usuário está bem clara na metamensagem dos signos metalinguísticos. Nos signos estáticos, esta identificação se deu a partir da percepção dos tipos de signos estáticos utilizados como *links* ou signos convencionais. Nos signos dinâmicos, a interação se dá a partir de interações simples de cliques ou toques nos signos estáticos para atingir os objetivos desejados, como obter mais informações a respeito de um determinado conteúdo. Um ponto que precisa ser destacado é que em alguns momentos não é explicado como o usuário deve interagir com a interface através do uso dos signos estáticos. O *designer* assume que o usuário é alguém experiente no uso de interfaces *web* ou *mobile*, e que os elementos utilizados (*links*, botões) façam parte do seu cotidiano. Dessa forma é necessário que o usuário explore a interface para conseguir perceber os detalhamentos de explicação, já que o *designer* não informa como ela deve ser usada.

Com relação à inconsistência, em alguns momentos não foi mantido o padrão nos signos estáticos, como mostra a Figura 8. Quando o usuário clica no primeiro tópico para obter mais informações a respeito dele, o ícone, que antes era uma maleta roxa, se transforma na

logomarca da empresa anunciante. No conteúdo apresentado a partir da interação com o menu "Saiba mais" (Figura 9), as informações apresentadas estão separadas por categorias: atividades nos Produtos Facebook (interesses, categorias, públicos semelhantes), atividades com outras empresas (públicos personalizados, atividade *off-line*) e informações de localização. Em nenhum momento o *designer* faz relação dessas categorias com os signos estáticos apresentados na interface de explicação simplificada (Figura 8), ficando a comunicação destes ícones incompleta.

Em boa parte do processo de comunicação do *designer*, ele tenta deixar clara a mensagem para o usuário. Esse fato é observado pela apresentação de signos estáticos associados a signos metalinguísticos. Ele consegue, através de redundância, explicar claramente boa parte dos signos utilizados na interface.

Como se trata de uma explicação a respeito de o porquê do usuário estar vendo determinado anúncio (signo dinâmico da interface), esse conteúdo já se trata na verdade de um grande signo metalinguístico. Em outras palavras, a explicação em si é um signo metalinguístico do anúncio a que se refere. Existe uma boa distribuição entre signos metalinguísticos e estáticos. Boa parte dos signos estáticos está acompanhada de signos metalinguísticos, porém devido à interface não prover muita interação ao usuário, para atingir seu objetivo de entender a exibição do anúncio, a interface comunica pouco através de signos dinâmicos, mesmo assim mantém um padrão de comunicação, exibindo mensagens na forma de textos simples para o usuário. A comunicação por signos dinâmicos está no processo de detalhamento de explicações e palavras-chave ao interagir com a interface (Figura 10).

O *designer* utilizou uma estratégia textual para explicar por que o usuário está vendo o anúncio. Estas explicações podem ser detalhadas a partir da interação do usuário sempre que ele necessitar obter mais informações a respeito de determinado tópico ou informação. Os signos estáticos incluídos da explicação em sua maioria são exibidos com signos metalinguísticos para uma compreensão rápida da interface. A linguagem utilizada na comunicação é simples, e remete a um diálogo direto do *designer* para o usuário. Existe um pequeno ruído na comunicação quando o *designer* utiliza alguns signos estáticos para categorizar as regras utilizadas na recomendação de publicidades, porém ele não explica em nenhum local qual o significados destes signos, como : maleta, silhueta do usuário e símbolo do *Facebook* (Figura 8).

A partir do MIS concluímos que o *designer* conseguiu comunicar bem a explicação aos seus usuários-alvo, podendo abranger usuários com perfis variados. Esse fato se comprova a partir de soluções como utilização de linguagem textual simples, signos estáticos acompanhados

de signos metalinguísticos para facilitar o entendimento do usuário, explicações de forma gradual, em que o usuário tem a opção de ver uma explicação simplificada ou se aprofundar no assunto tendo uma explicação mais completa a respeito dos motivos de estar vendo algumas publicidades. Algumas rupturas de comunicação foram observadas, devido à falta de padronização na utilização de alguns signos estáticos e, por vezes, à existência de alguns signos que não têm significado explícito para o usuário, o que nos leva a refletir a respeito da necessidade de existência deles, ou mesmo referenciá-los quando são explicados de forma detalhada dentro do menu "Saiba mais". A padronização de alguns textos utilizados na explicação, modificando somente o conteúdo que diz respeito aos usuários e apresentados em negrito, torna a explicação simplificada muito vaga e genérica, não ficando claros os reais motivos da recomendação feita. Não existe uma explicação em relação a como interagir com a interface, ficando essa *expertise* a cargo da experiência dos usuários no uso de aplicativos *mobile* e navegação em páginas de internet, já que ele baseou toda a navegação e usabilidade nessa modalidade de uso. Outro ponto importante a ser destacado está no fato de o *designer* não tentar explicar, mesmo de forma simples, como o algoritmo de recomendação realmente funciona, embora, no menu "Saiba mais"(Figura 9), mesmo de forma superficial, exista uma explicação, como funcionam as técnicas de filtragem utilizadas para inferência das recomendações.

5.2.2 Resultados do questionário

O questionário ficou disponível de 11 a 15 de setembro de 2019 e foi compartilhado em grupos de Whatsapp e Facebook. Foram obtidas um total de 63 respostas, sendo excluídos da amostra 5 questionários (4 relacionados a usuários que informaram não usar o *Facebook* e 1 duplicado). Ao final ficamos com uma amostra de 58 questionários válidos para a análise de dados, dos quais:

- **Sexo:** Masculino: 41 (70,7%) / Feminino: 17 (29,3%);
- **Idade:** 18-29: 18 (31,0%) / 30-39: 25 (43,1%) / 40-49: 12 (20,7%) / Acima de 50: 03 (5,2%);
- **Grau de Instrução:** Ensino Médio: 9 (15,5%) / Ensino Superior: 25 (43,1%) / Pós-Graduação: 24 (41,4%).

Das pessoas que responderam, 31 (53,4%) trabalham em outras áreas que não têm relação com tecnologia, os 27 restante (46,6%) têm algum envolvimento com TI, seja estudando ou trabalhando. Quanto ao uso do Facebook, 50 respondentes (86,2%) informaram acessar pelo

smartphone, o restante (08 - 13,8%) acessa pelo computador. Quanto à frequência de acesso, 40 (69,0%) acessam mais de 3 vezes por semana, 11 (19,0%) acessam de 2 a 3 vezes por semana e 7 (12,0%) acessam só uma vez por semana.

Quando perguntados, em uma escala de 1 a 8, qual o grau de relação das publicidades apresentadas com o perfil dos usuários: (56,9%) informou grau 5 ou 6 para essa relação; 20,7% atribuíram grau 7 ou 8; 17,3% atribuíram grau 3 ou 4; e somente 5,1% informação grau 1 ou 2.

Quanto ao interesse em entender por que determinadas publicidades aparecem em seus (usuários) *feeds*: 50,8% dos respondentes informaram grau 7 ou 8; 29,3% informaram um grau de interesse 5 ou 6; 6,9%, grau 3 ou 4 e 12% informaram grau 1 ou 2. Embora a maioria tenha demonstrado interesse nas explicações, 69,0% dos respondentes nunca haviam acessado o menu "por que estou vendo esse anúncio" disponibilizado pelo Facebook.

Quanto ao grau de entendimento da explicação fornecida no momento da realização das tarefas: 48,3% informaram ter um grau de entendimento 7 ou 8; 32,8% informaram grau 6 ou 7; 15,5% informaram grau 3 ou 4; e 3,4% informaram grau 1 ou 2 de entendimento da explicação. Embora boa parte tenha demonstrado ter um certo entendimento sobre as explicações apresentadas, quando perguntados se ainda tinham dúvidas sobre a explicação: 26 (44,8%) relataram apresentar algum dúvida; 24 (41,4%) deixaram a resposta em branco; e 8 (13,8%) responderam não ter dúvidas em relação à explicação fornecida. As dúvidas apresentadas estão descritas conforme os códigos a seguir:

- **Dúvidas sobre como foi feita a recomendação (8):** os usuários fazem questionamentos em relação ao funcionamento do algoritmo, ou a partir de que dados eles chegaram à conclusão de que aquela recomendação era relevante para eles. Ex.: *P15:"Eu não bebo, então fico me perguntando porque o Facebook me sugere anúncio de uma marca de cerveja."*.
- **Como os dados são obtidos/repassados ao anunciante (7):** os usuários têm uma preocupação com a privacidade dos dados utilizados pelo Facebook para fazer a recomendação. Ex.: *P21:"Como os dados foram obtidos ou que tipos de dados são disponibilizados pelo Facebook para seus anunciantes e como os algoritmos funcionam"*.
- **Consideram a explicação vaga (7):** para os usuários as explicações são muito genéricas e, com base no que é explicado, consideram que não é possível fazer recomendações precisas, a partir daquele conteúdo. Ex.: *P17: "As explicações dadas pelo Facebook são tão simples e genéricas de modo que não faz sentido haver publicidades tão específicas*

seguindo apenas essas explicações.”.

- **Conteúdo apresentado não corresponde ao perfil (3):** alguns usuários questionam a relevância de determinada publicidade para o seu perfil. Ex.: *P36:"não concordo com algumas das classificações que eles me deram, como por exemplo, Viajante frequente, pois raramente viajo mais do que uma vez por ano para algum canto fora do estado.”*.
- **Significado dos ícones (3):** alguns questionaram os significados dos ícones da interface.

Um respondente tinha dúvida em relação à frequência com que as publicidades aparecem, e um outro informou desinteresse no entendimento dessa explicação. Algumas respostas foram codificadas em mais de um categoria, não tendo assim uma categorização exclusiva.

Quanto ao entendimento dos usuários em relação aos textos exibidos em negrito na explicação, um total de 38 (65,5%) afirmaram compreender os motivos pelo qual os textos estavam destacados e 20 (34,5%) informaram não ter entendimento. As causas informadas, pelos usuários, para o texto em negrito, são descritas a seguir:

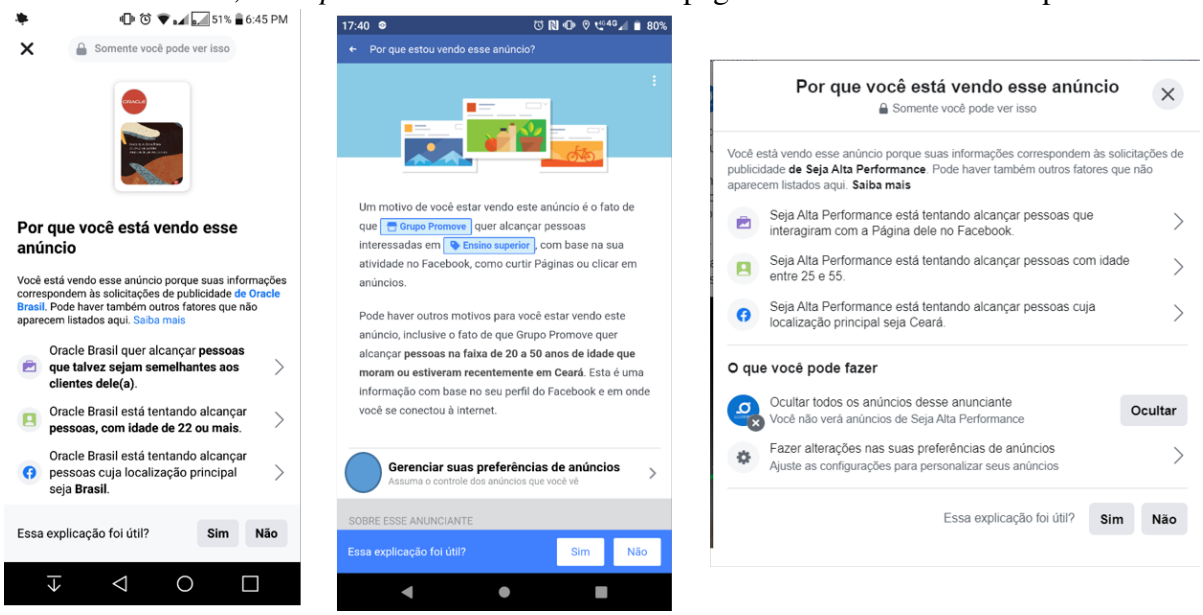
- **Destacar informação / Palavras-chave (39):** consideram que os textos exibidos em negrito são para chamar atenção para informações mais relevantes ou palavras-chave. Ex.: *P22: “Informação mais relevante”*, *P27: “Para ter mais atenção, destacar.”*.
- **Características do perfil desejado (10):** consideram que os textos exibidos em negrito têm relação com as características do perfil desejado. Ex.: *P5: “São os principais filtros de busca que facilitam a compreensão do algoritmo e o alcance do público e parâmetros requeridos.”*, *P43: “Chamar atenção pois o texto sempre eh o mesmo e o negrito eh a variação q se.enquadra a meu perfil”*.
- ***Link* para detalhar informação (1):** consideram que os textos exibidos em negrito são *links* que podem dar mais detalhes a respeito daquela palavra destacada. Ex.: *P17: “Porquê são links que levam a mais informações sobre relativas a palavra em negrito.”*.

Ainda nas respostas à pergunta anterior, 4 usuários deram respostas que não apresentam relação com a pergunta, 2 informaram que não sabem o que significam os textos em negrito e 1 usuário relatou não receber uma explicação com detalhes de textos em negrito.

Quando perguntados sobre o entendimento em relação aos ícones exibidos na tela de explicação (Figura 8): 33 (56,9%) respondentes informaram compreender o significado dos ícones; e 25 (43,1 %) informaram não entender o que significavam. Foi perguntado o que eles achavam que representam os ícones apresentados. Para esta pergunta, foram levados em

consideração somente os usuários que enviaram imagens de sua tela de explicação, pois era a forma que tínhamos de fazer um comparativo entre a resposta e a imagem. Foram recebidos um total de 23 imagens, dentre a quais 9 imagens de *smartphone* com aplicativo padrão do Facebook, 3 imagens do aplicativo Facebook Lite e 8 imagens do Facebook acessado via computador. Foram descartadas 3 imagens por não terem relação com a pesquisa. Para o estudo, desconsideramos as imagens do Facebook Lite, pois não tinha relação com o padrão das outras interfaces, conforme apresentado na Figura 11. Somente um usuário fez correspondência completa do ícones com sua real representação. Ex.: *P5:"Negócio (maleta), pessoal (ícone de contato), informações compartilhadas no Facebook (logo do Facebook)"*. A maioria dos usuários fez relação parcial dos ícones com seu real significado. Ex.: *P15: "Representam usuários com perfil parecido com o meu"*, *P41: "Talvez os dois primeiros se relacionam a filtros relacionados a qualidades específicas de "pessoas", e o que tem o logo do Facebook seja devido a relação dos usuários Brasil na plataforma, um atributo de localização, e não de característica das pessoas."*. Algumas respostas demostravam que os usuários não apresentavam nenhum entendimento a respeito dos ícones por exemplo: *P33: "Não tenho idéia"* ou *P47: "Não consegui identificar um padrão"*.

Figura 11 – Exemplos de interfaces enviadas pelos usuários a partir de *smartphone* com Facebook normal, *smartphone* com Facebook Lite e página do Facebook no computador



Fonte: Captura de tela de explicações sobre recomendações de publicidade do Facebook eviadas pelos participantes da pesquisa

.

Os usuários avaliaram as explicações em sua maioria com nota 3 (32,8%) ou 4 (39,7%). Mesmo apresentando alguns indícios de que não entenderam a explicação completa-

mente ou não se aprofundaram na explicação, a maioria dos usuários deu uma nota razoável para o conteúdo explicado.

A partir das imagens enviadas pelos usuários, percebemos que a interface de explicação não apresenta uma padronização entre as plataformas do *smartphone* e computador. Alguns ícones apresentados nas telas, os destaque de textos em negrito, ou até mesmo a disposição das informações diferem nas duas plataformas. A partir das respostas dos questionários, percebemos que os usuários, independente do perfil, têm entendimento parcial quanto aos ícones e aos seus significados, sentindo dificuldade principalmente no signo representado por uma maleta, o qual somente um usuário informou fazer relação com a categoria correta utilizada para recomendar itens. Quanto aos textos grifados em negrito, a maioria dos usuários deu significado superficial, definindo-os apenas como uma forma de destacar uma informação. As respostas relacionadas às dúvidas sobre as explicações se mantiveram balanceadas entre os dois tipos de usuários, sendo 14 usuários da área de tecnologia e 12 que não tinham relação com a área. Um usuário da área de tecnologia teve interesse em conhecer como os algoritmos de recomendação funcionam, no entanto, embora as explicações abordem os dados que são utilizados para realizar a filtragem das recomendações, não existe um aprofundamento a este nível. Não foi possível, através do questionário, saber se os respondentes que não entenderam as explicações avançaram até a tela de explicação detalhadas ou se ficaram somente na tela de explicação simplificada. A partir das respostas foi percebida uma preocupação dos usuários com a privacidade dos dados. Acreditamos que, quando criaram a conta no Facebook, não leram as políticas de privacidade dos dados a qual informa como suas informações serão usadas pelo Facebook ou não recordam o que consta nestes termos.

5.2.3 Resultados do experimento

Neste estudo utilizamos a Engenharia Semiótica para a analisar a qualidade da comunicação das explicações em recomendações de publicidade do Facebook. Uma das contribuições deste trabalho é o resultado desta análise, no qual apresentamos, a partir do MIS, uma visão do *designer* relacionada à emissão da mensagem contida na explicação e do questionário que nos permitiu ter uma visão de como os usuários recebem essa mensagem. Identificamos algumas boas práticas para comunicação desta explicação, como a utilização de textos simples e explicações de forma gradual dessas explicações aos usuários, que podem se aprofundar de acordo com sua necessidade, não comprometendo a interface com uma grande quantidade de informações.

Apesar das boa práticas, conseguimos elencar alguns problemas que podem causar rupturas de comunicação da explicação. A Tabela 4 apresenta os problemas encontrados com a informação de qual/quais métodos os registraram. O questionário, nos ajudou a validar algumas conclusões encontradas com o MIS, a partir da triangulação dos dados obtidos, mas devido as restrições em sua aplicação, não foi possível perceber se os usuários acessaram as explicações detalhadas da interface.

Tabela 4 – Problemas encontrados a partir da aplicação do MIS e do questionário

Problema	MIS	Questionário
Falta de explicações a nível de algoritmos	Sim	Sim
Falta de padronização de signos	Sim	Sim
Existência de signos estáticos que não são referenciados nos signos metalinguísticos	Sim	Não
Padronização de alguns textos da explicação, levando a uma explicação genérica e algumas vezes sem muito significado para o usuário	Sim	Sim

Fonte: Elaborada pelo autor
.

Uma outra contribuição deste estudo foi a exploração do uso da Engenharia Semiótica para avaliar explicações. Ela nos trouxe o conceito de comunicabilidade e balizou a escolha dos métodos de avaliação. Os *designers* de Sistemas de Recomendação (SR) podem, além da utilização de métricas centradas no usuários, utilizar a comunicabilidade como um objetivo a ser alcançado e que pode estar contribuindo para um melhor entendimento das explicações quanto a recomendações de itens.

Este trabalho não foca na engenharia de explicação em IA, mas usa a lente da Engenharia Semiótica para analisar a qualidade das explicações geradas. Estamos fazendo um exercício exploratório sobre como é analisar a qualidade das explicações através de uma engenharia de signos estáticos, dinâmicos e metalinguísticos. Um dos benefícios disso foi trazer para a avaliação o conceito de comunicabilidade. Em geral, as explicações em SRs são construídas e analisadas utilizando abordagens centradas no usuário e buscam atingir objetivos como transparência, escrutínio, confiança, persuasão, eficácia, eficiência e satisfação do usuário (TINTAREV; MASTHOFF, 2015).

Percebemos, pelo MIS, que há uma intenção do *designer* em explicar como foi feita a recomendação de publicidade, entretanto, quando perguntamos aos usuários, através do

questionário, vimos que ficaram algumas lacunas na compreensão das explicações relacionadas à privacidade dos dados, à coleta e ao repasse dos dados às empresas parceiras, e às informações sobre como os algoritmos funcionam. Percebemos também que as interfaces de explicação no Facebook, são personalizadas para o perfil do usuário e variam de acordo com a empresa anunciante. Assim, é importante que, no planejamento do MIS, essa característica seja levada em consideração, e a inspeção seja realizada em mais de uma publicidade ou por mais de um avaliador.

A utilização do questionário, nos permitiu coletar a opinião dos usuários e triangular os dados com os resultados do MIS. Apesar de conseguirmos bons resultados na combinação dos métodos, MIS e Questionário, umas das limitações da pesquisa foi a falta da aplicação de testes com usuários, pois não foi possível investigar como os usuários se aprofundaram nas explicações. A aplicação de técnicas como entrevistas e *think aloud*, associados ao MISI (Método de Inspeção Semiótica Intermediado) (OLIVEIRA; PRATES, 2018), acreditamos ser um caminho para sanar as lacunas deixadas e que poderão ser aplicadas em trabalhos futuros. Embora não tenhamos utilizado o MAC, para avaliar como os usuários recebem as explicações através da interface, por entender que uma interface de explicação composta, em sua maior parte, por signos metalinguísticos e poucas interações, ficaria difícil identificar possíveis rupturas na comunicação, consideramos importante a realização de um estudo com utilização desta técnica para verificação desta suposição, e uma reflexão sobre como ela pode contribuir na avaliação de interfaces de explicação.

O relatório deste experimento pode ser encontrado no seguinte endereço eletrônico: encurtador.com.br/luL46.

5.3 Definição do MoReXAI

5.3.1 Identificação dos pontos de discussão

Vários mapeamentos de princípios éticos foram propostos na literatura. Os objetivos destes mapeamentos é definir eixos principais e convergências a tópicos de princípios éticos gerais. Fjeld *et al.* (2020) fazem a leitura de 36 documentos de princípios éticos, dos mais variados contextos, e os mapeia em 8 dimensões: privacidade, responsabilidade, segurança e confiabilidade, transparência e explicabilidade, justiça (equidade) e não discriminação, controle humano da tecnologia, responsabilidade profissional e promoção de valores humanos. Este

conjunto de temas principais depois são reagrupados por Toreini *et al.* (2020b) em conjuntos de princípios éticos para IA confiável em 4 dimensões: Justiça, Explicabilidade, Responsabilidade e Segurança.

Em nosso trabalho tomamos como base a leitura do mapeamento de Burle e Cortiz (2019), realizado a partir da leitura de princípios éticos para IA em seis iniciativas internacionais, dentre elas: duas do setor governamental (Comissão Europeia (High-Level Expert Group on AI, 2019) e Departamento de Defesa Norte-americano (BOARD, 2019)), duas do setor empresarial (*Google* (GOOGLE, 2018) e *Microsoft* (MICROSOFT, 2019)), uma organização internacional (Organização para a Cooperação e Desenvolvimento Econômico - OCDE (OECD, 2019)) e outra composta de academia e setor empresarial (Academia de Inteligência Artificial de Pequim (PEQUIM, 2019)), no qual cria uma categorização dos princípios dividindo-os em seis dimensões principais: Privacidade e Segurança, Equidade, Confiabilidade e Segurança, Impacto Social, Responsabilidade, e Transparência. Além desse mapeamento, realizamos ainda a leitura de mais quatro documentos sendo três relacionados à organização da sociedade civil e das associações *multistakeholders* (*Uni Global Union* (UNION, 2017), *Institute of Electrical and Electronic Engineers - IEEE* (AUTONOMOUS; SYSTEMS, 2019), *ACM Committee on Professional Ethics* (ETHICS, 2018) e *Asilomar AI Principles* (INSTITUTE, 2017)), e uma do setor empresarial (IBM, 2019).

A partir destas leituras foram selecionados tópicos que apresentam relação com explicabilidade, tomando como base os objetivos dos usuários, conforme descritos a seguir:

– **Privacidade e Controle Humano (T1):** este tópico tem relação com a ideia de que os sistemas de IA devem respeitar a privacidade dos indivíduos no uso dos dados nesses sistemas, bem como fornecer às pessoas afetadas agência sobre seus dados e decisões tomadas com eles. A explicabilidade auxilia os usuários na compreensão de que dados são utilizados e qual peso eles exercem na tomada de decisão. Estão nesse princípio questões como controle para o uso dos dados por parte dos usuários, solicitação de consentimento para seja realizado tratamento nos dados, revogação do consentimento restringindo o uso quando solicitado, direitos de retificação e apagamento. Além disso acrescentamos a ideia de segurança, tendo em vista que a privacidade dos dados dos usuários pode vir a ser afetada em casos de uso indevido por parte dos desenvolvedores ou mesmo de terceiros. Os conjuntos de dados utilizados nos modelos devem apresentar medidas de segurança quanto a dados sensíveis de forma que os os dados dos usuários estejam protegidos e

anonimizados.

- **Responsabilidade e Prestação de Contas (T2):** este tópico tem relação com os princípios relativos à importância de mecanismos para garantir que a responsabilidade pelos impactos dos sistemas de IA seja distribuída de maneira apropriada, e que soluções adequadas sejam fornecidas, reconhecendo o papel vital que os indivíduos envolvidos no desenvolvimento e na implantação de sistemas de IA desempenham nos impactos dos sistemas. Nesse sentido, a explicabilidade se relaciona com esse principio, pois traz orientações relacionadas à avaliação de impacto dos sistemas inteligentes, bem como buscar formas com que esses sistemas possam ser auditáveis.
- **Cofiabilidade e Segurança (T3):** esse tópico tem relação com os requisitos para que os sistemas de IA sejam seguros, funcionando como pretendido (confiabilidade), e também protegidos e resistentes a acesso por terceiros não autorizados. Requerem que decisões importantes permaneçam sujeitas à revisão humana. Articula a ideia de que a explicabilidade dos sistemas inteligentes permitam a previsibilidade de suas saídas, como uma garantia de que não foram invadidos por terceiros.
- **Transparência e explicabilidade (T4):** esse tópico tem relação com a ideia de que os sistemas de IA sejam projetados e implementados para permitir a supervisão, inclusive por meio da tradução de suas operações em resultados inteligíveis e o fornecimento de informações sobre onde, quando e como estão sendo usados. A explicabilidade é importante para sistemas que podem causar danos, ou ter um efeito significativo sobre os indivíduos, impactando na qualidade de vida ou reputação de uma pessoa. Ela permite que os indivíduos que se sentirem afetados por um sistema de IA devam ser capazes de contestá-lo a partir de explicações fáceis de entender sobre os fatores e a lógica que serviu de base para a previsão, recomendação ou decisão.
- **Justiça, equidade e não discriminação (T5):** esse tópico tem relação com a exigência de que os sistemas de IA sejam projetados e usados para maximizar a justiça e promover a inclusão. Esse princípio articula que o preconceito em IA - nos dados de treinamento, nas escolhas do projeto técnico ou na implantação da tecnologia - deve ser mitigado para evitar impactos discriminatórios.

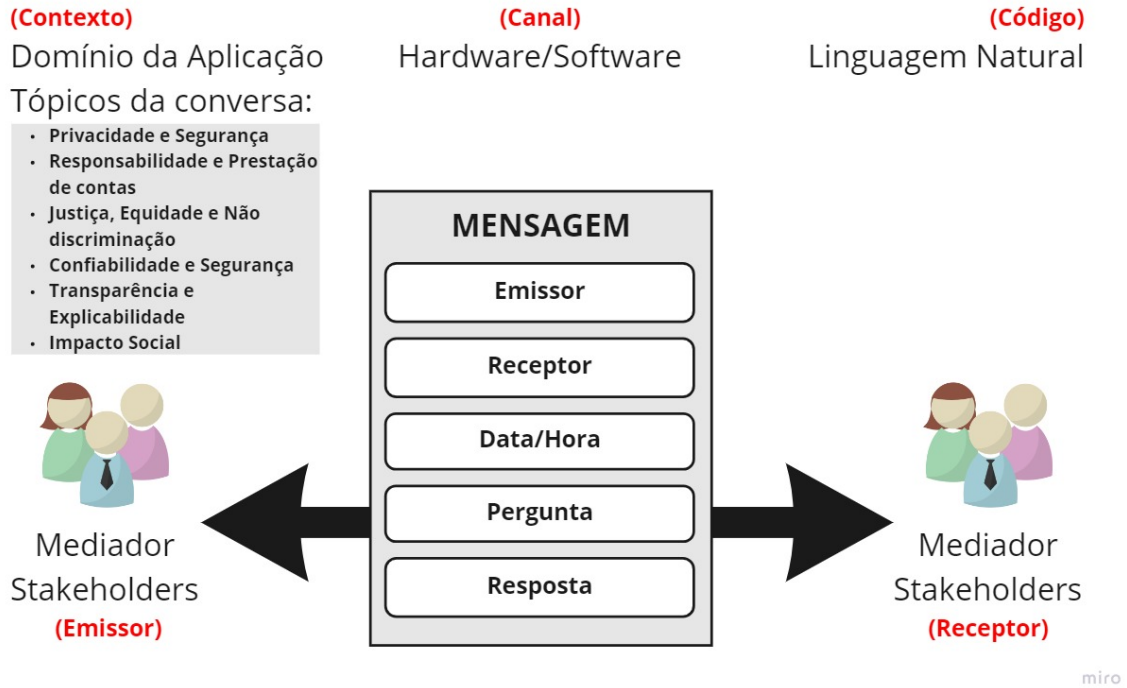
Após a definição dos tópicos de discussão, mapeamos os conjuntos de perguntas abordadas nas folhas de dados propostas em Gebru *et al.* (2018) para a etapa de desenvolvimento centrada em dados, conforme Tabela 5, e, para a etapa centrada no modelo, mapeamos as

perguntas relacionadas à estrutura dos cartões-modelo propostas em Mitchell *et al.* (2019), e os questionamentos sugeridos por Brandão *et al.* (2019), conforme Tabela 6. Ao lado de cada pergunta, apresentamos os tópicos de discussão ao qual elas estão relacionadas.

5.3.2 Formalização do MoReXAI

O modelo conceitual proposto nesta pesquisa objetiva estruturar a comunicação dos *stakeholders* de forma que eles reflitam a respeito de princípios éticos com o objetivo de extrair requisitos de *design* de explicações durante o processo de desenvolvimento de modelos de aprendizagem de máquina. Ele foi fundamentado no modelo de comunicação de Jakobson (1960) apud Santee e Temer (2011), possuindo assim todos os elementos propostos anteriormente por ele, como interlocutores (emissores e receptores), canal, mensagem, código e contexto. A Figura 12 traz uma representação do modelo e, a seguir, cada elemento é especificado.

Figura 12 – Modelo conceitual para raciocinar sobre *design* de explicações em sistemas de IA



Fonte: Elaborada pelo autor.

5.3.2.1 Contexto

O contexto envolve aspectos relacionados ao domínio da aplicação e os tópicos da conversa identificados na Seção 5.3.1 relacionados a princípios éticos. As conversas acontecerão

Tabela 5 – Perguntas mapeadas de Gebru *et al.* (2018) para a etapa centrada em dados

	Pergunta	T1	T2	T3	T4	T5
P1	Para que propósito foi criado o conjunto de dados? Havia uma tarefa específica em mente? Havia uma lacuna específica que precisava ser preenchida?	X				
P2	O que representam as instâncias que compõem o conjunto de dados? Existem vários tipos de instâncias? As relações entre instâncias individuais são explicitadas (por exemplo, classificações de filmes dos usuários, *links* de redes sociais)?	X		X		X
P3	O conjunto de dados representa todas as instâncias, ou é uma amostra de um conjunto maior? Se é uma amostra, qual foi a estratégia de amostragem?			X		X
P4	Há alguma informação faltando em instâncias individuais? Que estratégia foi usada para balancear o conjunto de dados? Como foi validada essa estratégia? Existem erros, fontes de ruído ou redundâncias no conjunto de dados?		X	X		X
P5	Os conjuntos de dados permanecem constantes ou podem ser modificados ou excluídos ao longo do tempo?			X		
P6	O conjunto de dados está relacionado a pessoas? Contém dados que podem ser considerados confidenciais ou sensíveis? Contém dados que, se visualizados diretamente, podem ser ofensivos, insultuosos, ameaçadores ou podem causar ansiedade?	X	X			X
P7	O conjunto de dados identifica alguma subpopulação (por idade, gênero)? Se sim, como essas subpopulações são identificadas e como elas estão distribuídas no conjunto de dados?	X		X		X
P8	É possível identificar indivíduos, direta ou indiretamente, a partir do conjunto de dados?	X	X			
P9	Como foi feita a coleta de dados? Quais mecanismos ou procedimentos foram usados? Como esses mecanismos ou procedimentos foram validados? Quem estava envolvido no processo de coleta de dados?	X	X	X	X	X
P10	Os indivíduos em questão foram notificados sobre a coleta de dados? Eles consentiram com a coleta e uso de seus dados? Como o consentimento foi solicitado? Se o consentimento foi obtido, foram fornecidos mecanismos para revogar seu consentimento no futuro ou para determinados usos?	X	X			
P11	Durante qual período de tempo os dados foram coletados? Esse período de tempo corresponde ao período de criação dos dados associados às instâncias?	X		X		X
P12	Foram realizados processos de revisão ética (por exemplo, por um conselho de revisão institucional)?	X	X	X		
P13	Foi realizada uma análise do impacto potencial do conjunto de dados e seu uso sobre os titulares dos dados (por exemplo, uma análise do impacto da proteção de dados)?	X	X			X
P14	Há algo sobre a composição do conjunto de dados ou a maneira como ele foi coletado e pré-processado / limpo / rotulado que possa afetar usos futuros? Há algo que um futuro usuário deve saber para evitar usos que possam resultar em tratamento injusto de indivíduos ou grupos ou outros danos indesejáveis? Existe algo que um futuro usuário poderia fazer para mitigar esses danos indesejáveis?		X			X
P15	Existem tarefas para as quais o conjunto de dados não deve ser usado?		X			

Fonte: Elaborada pelo autor.

no contexto da domínio da aplicação e levarão em conta como o processo de desenvolvimento dos modelos de aprendizagem de máquina pode impactar princípios éticos previamente definidos no modelo, levando em consideração a coleta e pré-processamento dos dados, treino, teste e

Tabela 6 – Perguntas mapeadas de Mitchell *et al.* (2019) e Brandão *et al.* (2019) para a etapa centrada no modelo de ML

	Pergunta	T1	T2	T3	T4	T5
P16	Qual o tipo de modelo de ML está sendo desenvolvido? Poderia explicar como ele funciona?				X	
P17	Qual algoritmo utilizado para treinar o modelo? Qual o grau de interpretabilidade do algoritmo? Poderia explicar como ele funciona?				X	
P18	Qual a principal intenção de uso do modelo? Quem são os usuários pretendidos que serão atendidos pelo modelo? Algum outro cenário de uso fora desse escopo?	X		X		
P19	O modelo pode afetar de alguma forma grupos demográficos ou fenotípicos? Qual a influência que esses fatores podem ter no desempenho do modelo? Como é feita a avaliação da influência desses fatores no desempenho do modelo?		X	X	X	X
P20	Os instrumentos ou o ambiente de coleta dos conjuntos de dados pode influenciar de alguma forma o resultado do modelo? Qual a influência que esses fatores podem ter no desempenho do modelo? Como é feita a avaliação da influência desses fatores no desempenho do modelo?	X	X	X		X
P21	Quais métricas de desempenho do modelo estão sendo relatadas e por que foram selecionadas em vez de outras métricas de desempenho? Como elas são calculadas?			X		
P22	Quais conjuntos de dados foram usados para avaliar o modelo? Por que esses conjuntos de dados foram utilizados? Como os dados foram pré-processados para avaliação?	X	X	X		X
P23	São feitas análises quantitativas com relação a subgrupos populacionais desagregados do conjunto de dados? Como o modelo se saiu em relação a cada fator? como o modelo se saiu em relação à intersecção dos fatores avaliados?		X	X		X
P24	Quais riscos podem estar presentes no uso do modelo? Quais estratégias de mitigação de risco foram usadas durante o desenvolvimento do modelo? Há algum uso pretendido para o modelo que seja preocupante do ponto de vista ético?		X			
P25	O modelo é escalável? Como garantir que o modelo inicial treinado e avaliado, quando aplicado no contexto real dos usuários se manterá com os mesmos resultados obtidos anteriormente? O modelo pode ser transferido do contexto pretendido para outro?		X	X		
P26	Existe documentação relacionada aos conjuntos de dados? E ao modelo?		X			
P27	O que é explicabilidade? Qual o objetivo de uma explicação? Pra quem é essa explicação? Como ela se apresenta? Quando ela deve ser apresentada? Onde ela deve ser apresentada?				X	

Fonte: Elaborada pelo autor.

avaliação do modelo de ML.

5.3.2.2 Interlocutores

Consideramos emissores e receptores todos os *stakeholders*. Os interlocutores, nesse contexto, podem desempenhar papéis como: desenvolvedores, cientistas de dados, acadêmicos, usuários-finais, entre outros. É importante captar os papéis dos interlocutores na conversa, para que, na análise, seja possível identificar qual a perspectiva de cada *stakeholder* sobre o tópico

discutido.

Um outro interlocutor importante nesse processo é o mediador, pois é ele quem vai conduzir a conversa entre os *stakeholders*. O mediador começa a conversa a partir de um conjunto de perguntas que serão discutidas pelos interlocutores. Como se trata de uma conversa multidisciplinar, cabe ao mediador fazer intervenções, no sentido de verificar se o entendimento do que está sendo conversado é compreensível para todos e solicitar ao emissor que sejam buscadas novas formas de explicar o assunto que está sendo abordado.

O mediador também é responsável por garantir, durante a conversa, que todos participem, podendo direcionar a conversa para determinados participantes como uma forma de que todos possam expor suas opiniões sobre o assunto. Além disso ele também é responsável por fazer o registro das conversas e posteriormente a extração dos requisitos de explicação. Nesse sentido o mediador deve ser um profissional que conheça um pouco do processo de desenvolvimento de modelos de ML para realizar a mediação da conversa, como também ter conhecimentos sobre *design* de interação para realizar a extração dos requisitos de explicação.

5.3.2.3 Canal

O canal corresponde ao *hardware* e ao *software* onde a conversa será executada, podendo inclusive acontecer em formato de grupos focais presenciais. Como se trata de um modelo conceitual, ainda não foram definidas quais ferramentas darão melhor suporte à conversa dentro do modelo, levando em consideração que ela pode acontecer de forma presencial ou remota, sendo esta última em formato síncrono ou assíncrono.

5.3.2.4 Mensagem

A mensagem é o componente do modelo que contém o conteúdo comunicado. Ela está relacionada aos tópicos da conversa (princípios éticos) que são propostos no modelo. O conjunto de mensagens trocadas a partir das respostas das perguntas lançadas pelo mediador traz uma perspectiva multidisciplinar dos *stakeholders*, a respeito de como eles pensam os princípios éticos dentro do domínio da aplicação, trazendo para essa conversa formas de explicar as abordagens utilizadas de modo que sejam compreensíveis para os envolvidos no projeto. A mensagem possui a estrutura apresentada a seguir (Figura 12):

- **Emissor:** identificador da pessoa que está enviando a mensagem. Pode ser, por exemplo, o nome da pessoa que está falando e o seu papel dentro do processo de desenvolvimento

do modelo.

- **Receptor:** identificador da pessoa que está recebendo a mensagem. Esta mensagem pode ser direcionada para todos da discussão, ou para pessoas específicas.
- **Data/Hora:** data e hora do envio da mensagem.
- **Pergunta:** este elemento está relacionado à etapa de desenvolvimento do sistema, que tem relação com o contexto da aplicação e os tópico da conversa (princípios éticos). Elas são selecionadas a partir de um conjunto de perguntas apresentados na Seção 4.3. O mediador escolhe as perguntas de acordo com o domínio da aplicação que está sendo desenvolvida e a etapa de desenvolvimento. Nem todas as perguntas precisam ser discutidas, sendo obrigatórias somente as perguntas contidas em P27 e que devem ser respondidas ao final da aplicação do MoReXAI após as perguntas selecionadas anteriormente pelo mediador.
- **Resposta:** texto escrito em linguagem natural que corresponde à resposta de uma pergunta no contexto do modelo, ou mesmo questionamentos e respostas levantados pelos interlocutores no decorrer das conversas.

Por se tratar de um modelo conceitual, no qual o canal não foi especificado, a estrutura da mensagem prevê elementos que são importantes independente do canal utilizado. Assim, são levadas em consideração nas mensagens, os interlocutores envolvidos, e em que contexto de perguntas está sendo conversado. As informações contidas na estrutura da mensagem são importantes no processo de analise das conversas para extração de requisitos.

5.3.2.5 Código

O código utilizado nas mensagens é a Linguagem Natural. Não sendo necessário conhecimento de nenhum código específico para haver a comunicação no modelo.

5.3.3 Fluxo de aplicação do MoReXAI

O modelo proposto envolve três grandes etapas: planejamento, rodas de conversa e extração de requisitos de explicação. A seguir detalhamos cada uma dessas etapas:

- **Planejamento**
  - **Definição dos objetivos:** nesta atividade são definidos os objetivos da aplicação do MoReXAI. Os objetivos que guiarão a escolha das perguntas que serão utilizadas na condução das conversas.
  - **Verificação do *status* de desenvolvimento do sistema:** esta atividade consiste em

realizar uma consulta detalhada a respeito do contexto no qual a aplicação se insere, observando dados técnicos relacionados aos conjuntos de dados e tipos de algoritmos que serão utilizados no treinamento do modelo de aprendizagem de máquina. Essa verificação é importante para a seleção do conjunto de perguntas que será utilizado na conversa.

  - **Definição do roteiro de aplicação:** nesta atividade é definido o roteiro de aplicação, no qual são selecionadas as perguntas que serão utilizadas com base nos objetivos e nas informações levantadas na etapa anterior. Além das pergunta selecionadas, sempre fará parte do roteiro as perguntas contidas em P27 relacionadas a explicações. É definido nesta etapa o formato como as rodas de conversas serão realizadas, podendo ser de forma síncrona presencial ou remota, ou assíncrona utilizando alguma tecnologia para a troca de mensagens (canal) definido previamente. A quantidade de rodas de conversas de aplicação do modelo são definidas a critério do mediador de acordo com o *status* de desenvolvimento e da quantidade de perguntas selecionadas, podendo ser aplicadas mais de uma roda de conversa.
  - **Recrutamento e entrevista com os participantes:** nesta atividade o mediador realiza o recrutamento dos *stakeholders* e realiza entrevistas para entender o grau de envolvimento deles com a aplicação que está sendo desenvolvida. No caso dos usuários, investiga os conhecimentos que já possuem sobre a área de aprendizagem de máquina e se conhecem aplicações de mesmo domínio daquela que está sendo desenvolvida.

- **Rodas de conversas:** Nesta etapa o mediador realiza as rodas de conversas seguindo o roteiro, conforme planejado na etapa anterior. Sugere-se que, dependendo da quantidade de perguntas, sejam aplicadas no mínimo duas rodas de conversas, sendo uma com perguntas relacionadas ao dados que serão utilizados no contexto da aplicação desenvolvida e outra com perguntas que se relacionam com os modelos de aprendizagem de máquina. Nas rodas de conversas, o mediador inicia apresentando o contexto da aplicação, os objetivos buscados com as conversas e como será a rotina da conversa entre os *stakeholders*.
- **Extração dos requisitos de explicação:** esta etapa consiste em o mediador realizar análise do discurso dos participantes nas conversas identificando requisitos de explicação. A partir das repostas às perguntas contidas em P27, o mediador poderá obter requisitos como: O que, onde, quando e como explicar.

6 ESTUDO DE CASO

Realizamos um estudo de caso exploratório para observar o uso do modelo em um contexto real de desenvolvimento de um sistema de IA. O relatório completo do experimento pode ser visualizado em: encurtador.com.br/antN4.

6.1 Descrição do estudo de caso

O estudo foi realizado em um projeto de pesquisa para o desenvolvimento de um sistema de recomendação de serviços do Governo do Estado do Ceará, que funcionará dentro do aplicativo Ceará App. O Ceará App é um aplicativo do Governo do Estado do Ceará com o objetivo de ofertar os principais serviços do governo de forma rápida e concentrada em um único local de forma remota. Os usuários podem acessar serviços relacionados a órgãos como: Departamento de Trânsito (Detran-Ce), Companhia de Água e Esgoto do Ceará (Cagece), Secretaria da Fazenda (Sefaz-Ce), Secretaria de Educação do Estado do Ceará (Seduc-Ce), Secretaria da Saúde (Sesa-Ce), entre outros. Os serviços disponibilizados são mais variados, entre eles estão: plantão *on-line* para atendimento de profissionais de saúde 24h para tratar problemas de saúde em geral, agendamento de teste Covid-19 e vacinação, emissão de certidões negativa e de regularidade, requerimento da carteira de habilitação (CNH) e solicitação de 2ª via ou renovação da CNH, anúncio, busca e compra produtos da agricultura familiar, entre outros. O estudo de caso foi dividido em duas etapas: preparação e aplicação do modelo.

6.2 Planejamento do estudo de caso

O planejamento do estudo de caso seguiu o fluxo do aplicação do MoReXAI, conforme descrito a seguir:

- **Definição dos objetivos:** explorar o uso do modelo no contexto real de desenvolvimento de um sistema de IA e observar como o modelo pode contribuir para o *design* de explicações no contexto da aplicação a partir das reflexões sobre princípios éticos dentro do processo de desenvolvimento. Avaliar o conjunto de perguntas utilizadas no modelo e a relação delas com os princípios éticos. Avaliar também o processo de aplicação do modelo, levando em consideração os papeis dos interlocutores e a forma síncrona como o modelo foi aplicado.
- **Verificação do *status* de desenvolvimento da aplicação de IA:** Observou-se que o sistema de recomendação de serviços em desenvolvimento se encontra na etapa inicial, estando os

desenvolvedores verificando quais algoritmos seriam utilizados. Eles contavam com uma base de dados de acessos dos usuários aos serviços do governo, coletados em um período de três meses, com mais de 130.000 registros de dados. As características contidas na base de dados eram relacionadas data de acesso ao serviço, nome do serviço, identificador do usuário, modelo-marca do dispositivo, tipo de dispositivo (*smartphone*/*tablet*), plataforma do dispositivo (Android/iOS), versão do sistema operacional, versão do Ceará App no dia do acesso e versão do SDK.

– **Definição do roteiro de aplicação do modelo:** o estudo de caso foi planejado para ser realizado de forma síncrona a partir de reuniões no Google Meet. As perguntas de discussão foram selecionadas tomando como base o contexto da aplicação e o estágio de desenvolvimento que ela se encontra (Tabela 7). A conversa dentro do modelo foi dividia em duas etapas, uma utilizando perguntas com o foco nos conjuntos de dados, e a outra mais voltada para as etapas de treinamento, teste e avaliação do modelo de ML.

– **Recrutamento e entrevista com *stakholders*:** consistiu no recrutamento dos dois pesquisadores envolvidos no desenvolvimento do sistema de recomendação e três usuários do Ceará App. As entrevistas aconteceram de forma remota pelo Google Meet e tiveram duração em torno de 10 a 16 minutos. As entrevistas foram gravadas, mediante acordo com os *stakeholders*, após explicações dos objetivos da pesquisa e leitura do termo de consentimento. A partir das entrevistas, verificamos que os pesquisadores têm experiência na área de desenvolvimento de sistemas de aprendizagem de máquina, sendo um doutor em computação, e um aluno de graduação do curso de engenharia de *software*, os quais vamos nos referir como (D1) e (D2), respectivamente. Os desenvolvedores estão trabalhando juntos em todas as etapas do ciclo de desenvolvimento do sistema proposto. Os desenvolvedores afirmaram que, no desenvolvimento do modelo, não haviam para ainda pra pensar a respeito de explicações, ou mesmo sobre princípios éticos. Quanto aos usuários recrutados, um primeiro usuário, (U1), tem formação em Letras Português e mestrado em Políticas Públicas, e já fez uso do Ceará App para agendamento de exames e cadastro para vacinação de Covid-19, além de agendar segunda via de habilitação no Detran-Ce. Um segundo usuário, (U2), é estudante de Engenharia da Computação e usa menos o aplicativo Ceará App, tendo baixado somente para agendar teste de Covid-19. O terceiro usuário, (U3), possui formação em Sistemas de Informação e cursa mestrado em Computação, tendo já aplicado alguns testes no aplicativo Ceará App, possuindo assim

um conhecimento mais aprofundado desse aplicativo.

Tabela 7 – Conjunto de perguntas abordadas em cada reunião de acordo com o estágio do ciclo de desenvolvimento da aplicação

Estágio de Desenvolvimento	Perguntas
Centrado nos dados	(P1), (P2), (P3), (P4), (P5), (P6), (P7), (P8), (P9), (P10), (P11), (P12), (P26)
Centrado no Modelo de ML	(P16), (P17), (P18), (P21), (P24), (P26), (P27)

Fonte: Elaborada pelo autor.

Além do fluxo do MoReXAI, foi realizado um grupo focal com o objetivo de avaliar o modelo proposto nesta pesquisa. Na ocasião foram realizadas perguntas relacionadas à importância de refletir a respeito de princípios éticos no contexto de modelos de aprendizagem de máquina, e se essa reflexão pode influenciar no processo de desenvolvimento e uso desses sistemas e auxiliar no *design* de explicações. As perguntas do grupo focal estão descritas a seguir:

- Vocês consideram importante refletir a respeito de princípios éticos, no contexto de modelos de aprendizagem de máquina? Por quê?
- Vocês acham que refletir a respeito desses princípios éticos pode influenciar nas estratégias utilizadas no modelo de aprendizagem de máquina? Por quê?
- Você acha que o modelo de discussão que foi aplicado traz alguma contribuição para o processo de *design* de explicações? Qual?

6.3 Rodas de conversas

A aplicação do modelo seguiu as etapas definidas no planejamento e aconteceu em dois encontros. O primeiro encontro teve duração de 1h, e, no segundo, houve uma extensão do tempo, para 1h30min, pois na ocasião foi aplicado o grupo focal de avaliação do modelo. A aplicação foi iniciada pelo mediador que explicou os objetivos da pesquisa e como aconteceria a conversa utilizando o modelo para raciocinar sobre o *design* de explicações em IA. Nessa explicação foram citados os princípios éticos aos quais estariam implícitos nas perguntas levantadas no decorrer da conversa. Além do mediador, a conversa contou com a participação de um profissional de IHC, que auxiliou no processo de coleta de informações, anotando dados que achava relevantes no decorrer da conversa. Foi solicitado aos desenvolvedores, que explicassem o sistema de recomendação que estava sendo desenvolvido. Após essa introdução foi dado inicio à sequência de perguntas, conforme consta na Tabela 7, de acordo com cada encontro.

6.3.1 Roda de conversa centrada nos dados

Nesta etapa foi identificado que o conjunto de dados não é documentado e que os dados não foram coletados pelos desenvolvedores do sistema de recomendação. Os desenvolvedores não souberam responder se houve consentimento para coleta dos dados, ou se os usuários foram informados da possibilidade de uso desses dados para desenvolvimento do projeto em questão. Nesse momento houve uma interação do grupo relacionada ao processo de solicitação de consentimento para o uso dos dados e como ele deveria ser apresentado aos usuários. Na interação do grupo sobre esse tópico, todos concordaram que os termos de consentimento são bastante extensos e nada atrativos para leitura e que, muitas vezes, clicam na autorização como uma forma de conseguir usar o sistema ou a aplicação. Deram várias propostas sobre como esse consentimento pode ser coletado, por exemplo: U3: *"Se pudesse ter uma forma mais usual, assim uma conversa como se fosse um 'chatzinho' pedindo alguma confirmação, alguma coisa"*; U1: *"tem que ter o documento detalhado a meu ver, só que também pode ter uma forma de apresentação desse documento de forma resumida, de forma mais atrativa"*; D1: *"O Android já tem um negócio parecido com esse, onde aparece lá, se você quer permitir acesso a tal periférico, como a câmera, o arquivo etc, então acho que seria uma abordagem, por que é uma frase bem curta, bem direto e objetivo..."*); D2: *"...como a Apple está fazendo agora, você tem o controle sobre o que que você está compartilhando de forma bem clara, visível, sem muita enrolação. Então eu acho que isso é um diferencial, uma coisa muito bacana... eu sinto a diferença. Com esse aplicativo aqui eu quero compartilhar tal informação, com esse outro aqui, eu não quero, então eu decido, eu gerencio os meus dados."*.

Os desenvolvedores informaram que a base não utiliza dados sensíveis, e que não é possível de alguma forma reverter o processo de anonimização. (U1) considera importante o processo de anonimização dos dados dentro do modelo, e que deve ser garantido de alguma forma a revogação de uso dos dados de forma clara (*"Você usa uma aplicação e não quer mais que seus dados sejam utilizados. Se a lei te garante isso, você vai querer isso, então o sistema vai ter que dar um jeito pra resolver isso. Eu sei que tem as limitações técnicas mas eu fiquei pensando no todo como usuário"*). Segundo os desenvolvedores foi utilizado todo o conjunto de dados para treinar o modelo e que não existem dados faltantes. Quanto à atualização da base de dados e como isso poderia afetar o modelo, os desenvolvedores informaram que ainda não haviam pensado a respeito da periodicidade de retreino do modelo. Nessa etapa, as perguntas e respostas foram muito técnicas, tendo pouca interação dos usuários, mesmo quando solicitados a

se pronunciarem quanto a algumas perguntas.

Os desenvolvedores informaram que como nas características da base dados não é possível identificar nenhuma subpopulação. Segundo eles foi pensando em criar um perfil econômico a partir dos modelos do aparelho acessado, porém não ficou definido isso até o momento da pesquisa. Foi então perguntado se é possível que o modelo criado possa automaticamente embutir um viés de classe a partir dessa característica. Segundo (D1) é possível, já que o modelo foi treinado com essa característica, ela exerce influência nas predições (*"Se você tem uma quantidade maior de pessoas que utilizam determinados serviços, então consequentemente a recomendação vai pegar perfis parecidos com aqueles usuários e vai recomendar os mesmos serviços..."*). Houve uma certa dificuldade por parte dos usuários no entendimento de como o viés pode ser embutido nos modelos de ML. Houve uma intervenção, por parte do mediador, para explicar esse processo. Quando perguntados a respeito de como mitigar esse viés, os desenvolvedores informaram que não foi pensado em nenhuma técnica que possa ser usada para mitigar o viés de classe identificado na conversa. Foi sugerido por (U2), que após o modelo treinado, fosse realizada uma avaliação e, à medida que for identificado viés discriminatório, fazer observações relacionadas a essa recomendação para que ela não aconteça novamente (*"... depois que a predição é feita, então faz-se uma análise nesse resultado da predição, então se for identificado alguma recomendação inadequada, ai essa recomendação é retirada..."*). De acordo com (U1), não existe interesse em aplicações do serviço público em embutir viés nos modelos de ML, tendo em vista que as aplicações não visam lucros (*"quando a gente trata de serviço público, não é um lucro que o serviço público ta tirando do seu acesso, ... eu não vejo muito isso de forma prejudicial no serviço público..."*).

6.3.2 Roda de conversa centrada no modelo de ML

O segundo encontro foi iniciado com uma explicação, por parte dos desenvolvedores, sobre o que é e como funciona um modelo preditivo. O desenvolvedor (D1) começou a explicar utilizando muitos termos técnicos, como multiplicação de matrizes e bibliotecas utilizadas nesse processo, porém depois exemplificou com recomendações de filmes e até do sistema de recomendação de serviços proposto. Falou inclusive da dificuldade de explicar sem a utilização de um recurso visual. Quanto aos algoritmos utilizados, os usuários consideram que não sentem necessidade de conhecer como funciona o algoritmo internamente, principalmente se ele acerta a recomendação, ou se não trabalha com dados sensíveis (U3: *"...pensar a fundo assim, querer*

saber, enquanto não mexer com dados sensíveis, eu acredito que não ia pensar".). Argumentam que gostam da praticidade que os sistemas de recomendação trazem, e que um explicação mais técnica não despertaria interesse na leitura (U1: *"Se tiver uma orientação geral talvez eu leia, mas se tiver algo muito técnico, mais aprofundado, eu como usuária que não sou da TI, eu acredito que não iria aprofundar a leitura da explicação não."*).

A conversa seguiu para um assunto relacionado ao enviesamento dentro de sistemas de recomendação, no qual (U1) citou que esse processo pode ser embutido de forma proposital como uma forma gerar lucros no setor privado, porém no setor público essa prática não é recorrente. Dentro do sistema de recomendação em questão, (D1) informou que ainda não foi criada nenhuma regra de negócio relacionado a esse processo de persuasão dos usuários, porém, se trabalhasse em uma empresa e essa regra de negócio é sugerida, considera que essa prática não é responsabilidade dele, mas de quem solicitou essa regra de negócio (*"...se vier um requisito que isso deve ser feito, acho que foge do controle do desenvolvedor e é mais uma discussão de negócio, porque afinal das contas o modelo ele vai recomendar o que você quiser, então você pode enviesar o modelo, ai entra outros problemas como de ética..."*).

Com relação às características que exercem mais influência na predição do modelo, (D1) informou que primeiramente faz uma análise humana, tentado identificar quais características são mais relevantes ou que podem ter um grande impacto na predição. Se o resultado não for bom, e tiver uma taxa de acerto menor que 90%, insere novas características no modelo, até que consiga fazer uma otimização e citou que já existem técnicas automatizadas para essa verificação. Os usuários informaram que consideram importante conhecer quais dados são utilizados para gerar as predições, segundo (U1) essa informação auxilia na escolha de continuar usando aquela aplicação, ou não (*"...acho importante, dependendo do tipo de dados que solicitam, eu nem continuo e acabo mudando de plataforma, vou para outro ambiente que não peça tantos dados meus"*). (U3) sugeriu que essa informação não fosse na forma de um texto muito extenso (*"Então se a gente tivesse mais ciência disso e não fosse daquela maneira mais burocrático de ler o contrato, eu acho que seria mais interessante para todo mundo querer saber né e ficar mais informado."*).

Quanto às métricas utilizadas para avaliar o modelo preditivo. (D1) citou um conjunto de termos técnicos relacionados às métricas e que ainda não tinham definido quais utilizariam no contexto do projeto. O mediador trouxe um exemplo da taxa de relevância da recomendação de filmes, apresentada pelo Netflix e consultou os usuários sobre a importância de conhecer as

métricas utilizadas na avaliação do modelo preditivo, ou algum recurso que desse esse indicativo, os usuários informaram que não buscam esse tipo de informação. O desenvolvedor (D2) citou que observa tanto no exemplo citado, como em *sites* de compra. A partir da fala de (D2), os usuários confirmaram que também fazem essas verificações.

A documentação relacionada ao desenvolvimento, está somente no contexto de controle de versão da evolução do modelo, a partir das alterações que vão sendo feitas. Os desenvolvedores informaram que não pensaram no uso de documentação que registre dados relacionados à responsabilidade e prestação de contas.

Ao final os usuários foram perguntados sobre os objetivos das explicações e o que, como, onde e quando explicar. Dente os objetivos das explicações foram citados o aumento da confiança (U1: *"...eu acho que é a questão da credibilidade, da confiança como eu acabei de falar pra você não achar que é apenas marketing..."*) e satisfação dos usuários (D1: *"Acho que se o usuário está satisfeito com aquilo, então é um dever para o sistema, fornecer essa alegria pra ele, seria uma fonte a mais de informação..."*). Quanto ao que deve ser explicado (U3) citou a forma como se chegou aquela recomendação (*"como foi chegado a esse resultado, com base em que, quais dados estão sendo usados pra lidar com isso"*). De acordo com (U1) a explicação deve ser prática e clara e informar os elementos utilizados para aquela recomendação (*"...eu imaginei indicadores, você falar quais são os elementos que compõe aquele sistema de recomendação, ... os elementos básicos que compõe aquele calculo de relevância e de recomendação. Acho que dá pra ter indicadores simples de entendimento simples, para que população entenda que foi a partir daquilo que gerou uma recomendação."*). Quanto à forma de exibição (U3) citou um formato de conversa (*"Acharia interessante, eu como usuário, ter tipo uma conversinha, como um bot, uma conversa com um robozinho, alguma animação, como se fosse uma conversa mesmo informal, poderia ser mais lúdico, até para usuários mais leigos ... e não que fosse uma leitura massante que eles na maioria das vezes não vão ler"*). Para (D2) as explicações devem ser apresentadas logo de cara e em uma linguagem simples e direta (*"tem que aparecer logo de cara para o usuário, no momento que ele inicializa o aplicativo e também estar disponível em alguma sessão do aplicativo caso ele queira curiosamente ir ler novamente. Acho que precisa ser de forma objetiva, clara, curta, simples. Falar qual dado está sendo usado, e que estão sendo usados para melhorar a experiencia dele na aplicação, e dessa forma a aplicação vai conseguir sugerir coisas que ele tem interesse, sem escrever demais. Eu como usuário não gosto de ficar lendo textão em aplicativo"*). (D1) apresentou várias sugestões para explicações das

recomendações (*"Mostrar quais características do usuário foram mais relevantes para chegar aquele resultado. Mostrar a data em que essa recomendação foi gerada para o usuário, por que pode ser uma recomendação muito antiga. Mostrar o perfil do usuário dentro desse sistema. Mostrar perfis similares..."*).

6.4 Extração dos requisitos de explicação

Nesta atividade realizamos a extração de requisitos de explicações a partir das falas no decorrer das conversas e das perguntas obrigatórias (P27) do modelo.

6.4.1 Por que explicar?

A pergunta "Por que explicar?"está diretamente relacionada aos objetivos buscados com as explicações. Os *stakeholders* citaram, no contexto do domínio da aplicação, que as explicações podem ter objetivos, como satisfação do usuário, passar confiança aos usuários, permitindo que estes tenham controle de seus dados, como também a necessidade de atender às regulamentações vigentes. A Tabela 8 apresenta os requisitos e trechos que falas que evidenciam este objetivo de explicação.

Observou-se que a melhoria da confiança dos usuários está relacionada ao entendimento de como o sistema funciona e quais dados são utilizados para geração das recomendações, bem como à ideia de que os usuários devem ter controle sobre os seus dados. Quanto ao atendimento das regulamentações vigentes, esse objetivo tem relação com a ideia de que o uso dos dados deve sempre estar disponível aos usuários, bem como a solicitação e revogação dos dados, esse objetivo também apresenta relação com os princípios de prestação de contas e responsabilidade.

6.4.2 O que explicar?

Quanto ao conteúdo que as explicações deveriam conter, houve várias sugestões. Uma delas foi explicar como a recomendação foi feita e com base em quais dados, ou mesmo por que o usuário está visualizando uma determinada recomendação. No caso de informar os dados utilizados para gerar a recomendação, enfatize aqueles que tiveram maior relevância nessa previsão. Nesse sentido, U1 disse: *"Pensei em indicadores (...) Os elementos básicos que compõem esse cálculo de relevância e recomendação. (...) se pegarmos o Spotify, o mais ouvido,*

Tabela 8 – Objetivos dos usuários quanto às explicações no contexto do estudo de caso

	Satisfação dos usuários
D1	*"Eu acho que você deve explicar porque é uma forma de você agradar ao usuário. Acho que se o usuário está satisfeito com aquilo então é o dever para o sistema fornecer essa alegria pra ele, seria uma fonte a mais de informação, cabe a ele se interessar ou não por aquela recomendação."*
	Melhoria na confiança dos usuários
U1	*"Eu também acho importante (informações sobre dados utilizados nas recomendações) dependendo do tipo de dados que solicitam eu nem contínuo. Acabo mudando de plataforma, vou para outro ambiente que não peça tantos dados meus (...) Esse tá invasivo demais, vou mudar. Mas saber os dados mais importantes, acho alguns tipos de aplicação você até percebe que dados seus eles estão utilizando para chegar aquela recomendação ali, não é normal quando é uma aplicação que você usa bastante."*
U1	*"Eu acho que é a questão da credibilidade, da confiança como eu acabei de falar, pra você não achar que é apenas marketing, a gente se sente muito vítima do marketing às vezes, recomenda-se aquele computador, aquela passagem, aquele filme, mas você não tem noção se aquilo é realmente uma recomendação de acordo com seu perfil de uso ou de acordo com o perfil de uso de todos os usuários da plataforma ou se é marketing, então acho que seria mais pra essa questão da confiabilidade."*
	Controle de uso dos dados por parte dos usuários
D2	*"Por fim eu acho que o grande lance é o controle que você tem sobre os seus dados, como a Apple está fazendo agora, você tem o controle sobre o que que você tá compartilhando de forma bem clara, visível sem muita enrolação. Então eu acho que isso é um diferencial, (...) e eu sinto a diferença, com esse aplicativo aqui eu quero compartilhar tal informação, com esse outro aqui, eu não quero, então eu decido, eu gerencio os meus dados".*
D1	*"uma questão muito do controle das informações, tem startups hoje em dia, se não me engano algumas localizadas nos estados unidos e na índia, que elas dão a opção para os usuários fornecer os dados deles e a partir disso monetizar esses dados, eles ficam sabendo o que está sendo feito com essas informações e ainda conseguem ganhar um dinheiro de volta podendo oferecer as informações deles, então acho que tudo gira muito em torno desse controle, de uma coisa que fica clara para o usuário, sem se tornar muito massante."*
	Atendimento a regulamentações vigentes
U3	*"(...)mas tem que ter tudo muito exposto por causa da lei, na lei exige que tenha tudo conforme o que vai ser feito, nos mínimos detalhes exposto, pra que o usuário tenha essa ciência(...)."*
U3	*"(...)o termo de consentimento tem que estar daquela forma, pela lei. Tem que estar tudo muito exposto."*
U1	*"O que eu imaginei quando vocês estavam falando, é que não é que a lei não obrigue ser o documento detalhado, eu acho que tem que ter o documento detalhado a meu ver só também pode ter uma forma de apresentação desse documento de forma resumida de forma mais Atrativa."*
U1	*" se a lei te garante isso, você vai querer isso, então o sistema vai ter que dar um jeito pra resolver isso, eu sei que tem as limitações técnicas, mas eu fiquei pensando assim no todo, como usuário, ele não percebe, ele não entende, como funciona esse sistema, essa base de dados, eu achei interessante quando vocês estavam explicando, mas realmente tem que haver formas de garantir isso, para quem entende e para quem não entende de desenvolvimento."*
D1	*"se o objetivo é realmente garantir que o usuário não vai ter os dados expostos, digamos assim, mas que de forma anônima é possível utilizar os dados, então se a lei permite isso, seria bem tranquilo se adequar a lei."*

Fonte: Elaborada pelo autor.

o mais baixado, o mais tocado. Acho que é possível ter indicadores simples e fáceis de entender, para que as pessoas que nem sabem que existe um indicador, mas entendam que foi a partir do que gerou uma recomendação".

Um ponto importante levantado é no caso de aplicativos que não são públicos, eles

devem informar aos usuários se a recomendação é algo relacionado ao *marketing*, se é uma recomendação patrocinada, e dizer por que e com base em que, estão recomendando aquele produto/serviço.

Também foi sugerido inserir na explicação em qual data a recomendada foi gerada dentro do modelo, porque pode, haver recomendações baseadas em dados antigos. Além disso, no caso de recomendações que consideram esses perfis semelhantes, discutiu-se a importância de explicar o perfil do grupo de usuários que está sendo utilizado para associar ao perfil da pessoa que recebe a recomendação.

Ainda sobre o que explicar, o modelo proporcionou uma conversa sobre a inserção de exemplos conhecidos dos usuários nas explicações. Nesse sentido o U3 disse: *"Sim, e semelhante a esse sistema de recomendação que já temos em alguns aplicativos, baseado no que sempre faço, mais vídeos vão aparecer, por exemplo no YouTube, semelhante ao que sempre vejo, por que eles vão me interessar, nesse caso o aplicativo Ceará será baseado no meu histórico de uso do aplicativo Ceará, sempre vai ter alguma função relacionada a esse uso, semelhante, que é o próximo passo, né?"*.

Notou-se que boa parte das respostas dos usuários, relacionadas ao que deve ser explicado, tinham relação com os princípios éticos de privacidade e justiça e não discriminação, como a preocupação com a anonimização dos dados, bem como saber quais dados estavam sendo utilizados na predição, ou até mesmo informações relacionadas a consentimento e revogação dos dados de forma mais clara. Essa abordagem tem relação com a confiança dos usuários.

No decorrer da conversa, surgiram várias situações que devem ser evitadas no conteúdo das explicações e que já são recomendações existentes na literatura como utilização de textos longos e termos técnicos. Além disso, citaram que não têm interesse em conhecer uma explicação mais aprofundada de como o algoritmo funciona, ou mesmo as métricas utilizadas para avaliar o modelo de predição, sendo enfatizado pelos desenvolvedores que alguns algoritmos são bem complexos de serem compreendidos. Foi citado pelos *stakeholders* que nem sempre a aplicação desperta o interesse por explicações, principalmente quando se tratam de contextos que não trazem nenhum prejuízo para os usuários.

6.4.3 Como explicar?

Falou-se sobre o projeto de explicação: os participantes reconhecem que as explicações não devem ser longas e devem ser apresentadas gradualmente, sempre que possível

usando exemplos, podendo também ser apresentadas em formas conversacionais, por exemplo U3 disse: *"eu acho, eu como um usuário, ter uma conversinha, tipo um bot, uma conversa com um robozinho, alguma animação, como se fosse uma conversa bem informal, poderia ser mais lúdico, mesmo para usuários leigos... e não que fosse chato lendo que na maioria das vezes não vão ler."*.

6.4.4 Onde explicar?

Foram sugeridos alguns locais onde deveriam aparecer as explicações, podendo vir junto com as recomendações, ficando em um local específico do aplicativo onde os usuários podem fazer essa consulta sempre que sentirem necessidade. O local onde as explicações deveriam ser apresentadas partiu de questões relacionadas aos termos de autorização para uso dos dados. Por exemplo, D1 disse: *"Acho que deveria ser mostrado junto com a recomendação, ter um link ao lado, entenda mais como essa recomendação foi gerada, por exemplo."* e D2 disse: *"(...) estar disponível em alguma sessão da aplicação caso queira, por curiosidade, voltar a lê-lo."*.

6.4.5 Quando explicar?

Também foi discutido quando explicar, os participantes chegaram a um consenso de que as explicações deveriam ser apresentadas próximas às recomendações (no caso de um sistema de recomendação). Nesse sentido, D2 disse: *"tem que aparecer imediatamente para o usuário, no momento em que ele iniciar o aplicativo e também estar disponível em alguma sessão do aplicativo caso ele queira lê-lo por curiosidade novamente."* e U3 disse: *"pode ser um passo a passo na inicialização do sistema, como um tutorial de como mover e dentro desse tutorial, está dizendo que esses dados serão usados, o que será usado e como será usado, e ficará guardado lá para quando ele quiser olhar, ou em alguma tela, né?"*.

A ideia de controles configuráveis para visualização de explicações também surgiu. Nesse sentido, U3 disse: *"(...) eu acho que deveria ser uma configuração, tipo, quando a gente vai fazer uma configuração, a gente permite ou não, por exemplo no celular, a gente permite ou não o uso de dados móveis, em determinada situação, acho que deveria estar no perfil, quando o usuário for ver a configuração. Porque por lei tem que estar à disposição do utilizador a qualquer momento, toda a informação (...)"*.

6.4.6 Para quem explicar?

Durante a conversa, ficou muito claro entre todos os participantes que as explicações no contexto do sistema de recomendação deveriam focar nos usuários que utilizam este sistema.

A Tabela 9 apresenta um resumo dos requisitos de explicação extraídos a partir da aplicação do MoReXAI. Mais detalhes sobre as falas dos *stakeholders* relacionados a cada um dos requisitos de explicação podem ser acessados no relatório contigo no seguinte endereço eletrônico: encurtador.com.br/antN4.

Tabela 9 – Requisitos de explicação extraídos a partir da aplicação do MoReXAI

Pergunta	**Requisito**
Por que explicar?	Satisfação dos usuários
	Melhoria da confiança dos usuários
	Controle de uso dos dados por parte dos usuários
	Atendimento a regulamentações vigentes
O que explicar?	Como as recomendações foram feitas e baseadas em que dados
	Informações sobre quando a recomendação foi feita
	Informações sobre o perfil dos usuários ou grupos de usuários
	Utilização de exemplos para gerar explicações
	Preocupação com o anonimato e quais dados foram coletados
	Não há interesse em informações sobre o funcionamento dos algoritmos
	Devem ser evitados termos técnicos
Como explicar?	Explicações por exemplos
	Comunicação das explicações em formatos diversos
	Explicações de forma gradual
	Utilização de linguagem simples
	Evitar uso de textos longos
	Evitar uso de termos técnicos
Onde explicar?	A explicação deve estar sempre disponível
	Explicações devem estar próximas às recomendações diversas
	Devem ter controles configuráveis de visualização das explicações
	Junto das recomendações ou em local disponível na interface
Quando explicar?	Evitar mensagens com solicitações excessivas aos usuários
	A explicação deve estar sempre disponível

Fonte: Elaborada pelo autor.

6.5 Análise de resultados da avaliação do modelo

Na avaliação do modelo realizada com o grupo focal final, os usuários trouxeram depoimentos bastante relevantes quanto às questões abordadas e como elas podem influenciar o processo de desenvolvimento deles, como observados nas falas de (D2): *"...de forma geral confesso que boa parte das perguntas expandiram um pouco a minha visão. Acho que uma falha que cometo muito como desenvolvedor é pensar somente como desenvolvedor, e esqueço que*

o que eu estou fazendo é para um usuário é para uma pessoa, então essas perguntas que você fez me ajudou bastante a refletir encima do que estou trabalhando e outras preocupações que precisa ter.", e na fala de (D1): *"acho que foi bastante relevante essas reuniões, muda a visão da gente sobre algumas coisas que geralmente já estavam engessadas, então eu vou alterar a forma como eu fazia as coisas a partir dessas reuniões, por que eu vou tentar facilitar, ou pelo menos tentar deixar um arcabouço de como fazer essas coisas que a gente estava conversando aqui, dado que eu concordo que são coisa boas.".* Do ponto de vista dos usuários, participar das discussões trouxe um novo olhar para os sistemas que utilizam (U1): *"achei super interessante, ainda não tinha parado pra pensar no tanto que os sistemas de recomendação estão presentes, nos aplicativos que uso. Achei super interessante esse viés da questão ética, geralmente se fala somente no sigilo de dados, mas o arcabouço ético, até chegar a esse dado, antes de ter esse dado, deveria ter uma preocupação ética para o uso. muitas vezes a gente se preocupa somente com final quando o dado já está lá e não tem essa preocupação ética antes."*

6.6 Discussão

6.6.1 Sobre o caráter epistêmico do MoReXAI

O estudo de caso realizado permitiu-nos refletir sobre os elementos do modelo bem como a sua utilização. Levar os desenvolvedores a refletir sobre o artefato que estão desenvolvendo junto com os usuários é bastante rico para construir explicações mais eficazes. Os desenvolvedores trouxeram depoimentos sobre as questões do modelo e como elas podem influenciar no seu processo de desenvolvimento, como observado nas falas de D2: *"...confesso que a maioria das questões ampliou um pouco a minha visão. Acho que um erro que cometo muito como desenvolvedor é pensar só como desenvolvedor, e esqueço que o que estou fazendo é para um usuário, é para uma pessoa, então essas perguntas que você fez me ajudaram muito a refletir o que estou trabalhando e outras preocupações que você precisa ter."*, e na fala de D1: *"Acho que esses encontros foram muito relevantes, muda a visão das pessoas sobre algumas coisas que normalmente já estavam no elenco, então vou mudar a forma como eu fazia as coisas com base nessas reuniões, porque eu vou tentar facilitar, ou pelo menos tentar deixar uma estrutura de como fazer essas coisas que a gente estava falando aqui, já que eu concordo que são coisas boas."*.

Além disso, o modelo trouxe um novo olhar aos usuários sobre os sistemas que

utilizam. Por exemplo, U1 disse: *"Achei super interessante, não tinha parado para pensar em quantos sistemas de recomendação estão presentes nos aplicativos que uso. Achei super interessante esse viés da questão ética, falando de modo geral apenas sobre a confidencialidade dos dados, mas o arcabouço ético, até chegar a esses dados, antes de ter esses dados, deve ter uma preocupação ética quanto ao seu uso. Muitas vezes a gente só se preocupa com o fim quando os dados já estão lá e não temos essa preocupação com a ética antes"*. Ainda sobre o caráter epistêmico do modelo, notamos que as explicações dadas pelos desenvolvedores aos usuários, em formato técnico, traziam termos que não são adequados para serem utilizados na explicação, como: *"...modelo preditivo, banco de dados, relacionamento de uma matriz..."*. Nesse momento, o mediador tinha o papel de ajudar a traduzir a explicação dos desenvolvedores para os usuários e também verificar com os usuários se ela foi compreendida. Entendemos que foi um momento rico para saber quais termos devem ou não ser utilizados nas explicações do sistema em desenvolvimento.

6.6.2 Melhorias no MoReXAI

Percebemos que algumas perguntas do modelo precisam ser melhor explicadas aos usuários. Uma maneira de fazer isso é usar exemplos gerais, preferencialmente de outro sistema conhecido do grupo. Por exemplo, no estudo de caso, algumas conversas ocorreram usando o sistema de recomendação de filmes da Netflix como exemplo. Portanto, percebemos que é importante orientar o mediador a acrescentar exemplos relacionados às perguntas e aos princípios éticos que elas representam na fase de planejamento.

Tínhamos previsto que o modelo seria mediado por alguém familiarizado com os elementos do modelo e que organizaria a conversa. Entre as funções que prevemos estão: definir o escopo da aplicação a ser discutida, definir e convidar os participantes da discussão, agendar e conduzir a discussão, analisar e compilar os dados coletados. No entanto, no estudo de caso percebemos que um papel importante do mediador é auxiliar na comunicação entre desenvolvedores e usuários. Durante o experimento, notamos que o desenvolvedor usou termos técnicos algumas vezes. Nesses casos, o mediador deve realizar uma "tradução" do que foi dito, ou mesmo intervir para que os desenvolvedores busquem outras formas de explicá-lo aos usuários, devendo o mediador acompanhar o que está sendo dito pelos técnicos e verificar se os usuários estão entendendo. Esse processo é interessante para capturar o sistema de significados compartilhado pelo grupo. A princípio, pensamos que o mediador era um especialista em

Interação Humano-Computador, mas percebemos que ele também precisa ter conhecimentos básicos sobre aprendizagem de máquina.

Quanto ao canal, como se trata de um modelo conceitual e não prevê uma ferramenta que dê suporte para as conversas, o estudo de caso aconteceu de forma síncrona através do Google Meet. Esse formato trouxe falas espontâneas dos *stakeholders*, que expressaram livremente seus conhecimentos e dúvidas, sem um tempo extra para refletir de forma aprofundada sobre o assunto conversado. Como trabalho futuro, sugerimos explorar o uso de uma ferramenta assíncrona para a troca de mensagens, na qual os envolvidos possam ter um tempo extra para pensar e estruturar melhor o raciocínio.

O estudo de caso foi realizado no contexto de uma aplicação em fase inicial de desenvolvimento. Uma das limitações do MoReXAI é que, embora tenhamos extraído requisitos de explicação a partir das conversas estruturadass no modelo, não foram instanciadas interfaces de explicação para discussão por parte dos *stakeholders*. Uma expansão do modelo poderia ter como parte do código a utilização de linguagem de interface, como acontece no modelo proposto por Sampaio (2010).

7 CONCLUSÕES E TRABALHOS FUTUROS

Propomos um modelo conceitual para apoiar a elicitação de explicações em projetos de ML. Usamos uma abordagem que envolve a participação do usuário e é baseada no *design* centrado na comunicação.

Concluímos que as declarações dos usuários relacionadas a tópicos de princípios éticos (privacidade, segurança, responsabilidade, confiabilidade, transparência, explicabilidade, justiça, equidade e não discriminação) geraram requisitos importantes para o *design* de explicações. O modelo promoveu a conversa sobre esses princípios e, em seguida, sugeriu ideias de explicações para a interface, dadas pelo próprio usuário. Foi possível falar sobre o que, por que, como, quando e quem explicar.

Além disso, percebemos o caráter epistêmico do modelo, pois todos os participantes da conversa afirmaram ter mudado sua visão sobre os pontos abordados. O estudo de caso nos trouxe a oportunidade de refletir sobre o uso síncrono ou assíncrono do modelo proposto. O fato de os dois encontros serem síncronos foi bastante rico, pois o contato entre os *stakeholders* permitiu maior envolvimento e engajamento na conversa. Além disso, o mediador teve a oportunidade de provocar os participantes para que todos participassem dando sua opinião. Por outro lado, a principal vantagem da conversa assíncrona é dar tempo para que as pessoas reflitam sobre as questões.

Embora o estudo de caso não tenha permitido mais tempo para os participantes refletirem sobre as questões da conversa, como as conversas eram síncronas, percebemos que o fato de termos uma entrevista dias antes de iniciar as conversas já levava os participantes a pensar sobre o que estavam falando. Além disso, como as conversas ocorreram em duas seções, houve tempo entre uma seção e outra, neste caso foram dois dias, para que os envolvidos refletissem sobre as questões.

Em estudos posteriores, pretendemos explorar o uso de uma ferramenta assíncrona, ou mesmo uma metodologia mista com momentos síncronos e assíncronos, que permita aos envolvidos ter tempo para refletir sobre as questões do modelo. Independentemente de as conversas serem síncronas ou assíncronas, o mediador terá o papel de manter o engajamento do grupo na conversa, por meio de mensagens direcionadas, garantindo a participação de todos.

Imaginamos o uso do modelo proposto em um cenário de construção de sistemas de IA em que há interesse em projetar explicações. Nesse contexto, vislumbramos um especialista em IHC interagindo com a equipe de IA para trabalhar em conjunto nesse desafio de projetar

explicações. Essa equipe convidará os usuários para participar das conversas. Essas conversas podem acontecer várias vezes, com usuários diferentes.

Nesse contexto, o modelo funciona como uma ferramenta epistêmica geradora de conhecimento, tendo em vista que, a cada aplicação dessa equipe, mesmo em contextos diferentes e variados, os interessados vão agregando conhecimento sobre o *design* de explicações.

REFERÊNCIAS

ANGWIN, J.; LARSON, J.; MATTU, S.; KIRCHNER, L. **Machine Bias**: There's software used across the country to predict future criminals and it's biased against blacks. 2016. Disponível em: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Acesso em: 22 abr. 2022.

ARRIETA, A. B.; DíAZ-RODRíGUEZ, N.; SER, J. D.; BENNETOT, A.; TABIK, S.; BARBADO, A.; GARCIA, S.; GIL-LOPEZ, S.; MOLINA, D.; BENJAMINS, R.; CHATILA, R.; HERRERA, F. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. [*s. n.*], [*S. l.*], v. 58, p. 82 – 115, 2020. ISSN 1566-2535. Disponível em: http://www.sciencedirect.com/science/article/pii/S1566253519308103. Acesso em: 22 jan. 2021.

AUTONOMOUS, T. I. G. I. on Ethics of; SYSTEMS, I. **Ethically Aligned Design**: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. 2019. Disponível em: https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html. Acesso em: 01 mai. 2022.

BARBOSA, C. M. de A.; PRATES, R. O.; SOUZA, C. S. de. Identifying potential social impact of collaborative systems at design time. In: BARANAUSKAS, C.; PALANQUE, P.; ABASCAL, J.; BARBOSA, S. D. J. (Ed.). **Human-Computer Interaction – INTERACT 2007**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 31–44. ISBN 978-3-540-74796-3.

BARBOSA, S.; SILVA, B. **Interação humano-computador**. [*S. l.*]: Elsevier Brasil, 2010.

BARBOSA, S. D. J.; BARBOSA, G. D. J.; SOUZA, C. S. d.; aO, C. F. L. A semiotics-based epistemic tool to reason about ethical issues in digital technology design and development. In: **Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency**. New York, NY, USA: Association for Computing Machinery, 2021. (FAccT '21), p. 363–374. ISBN 9781450383097. Disponível em: https://doi.org/10.1145/3442188.3445900. Acesso em: 22 jan. 2022.

BARBOSA, S. D. J.; PAULA, M. G. de. Designing and evaluating interaction as conversation: A modeling language based on semiotic engineering. In: JORGE, J. A.; NUNES, N. J.; CUNHA, J. Falcão e (Ed.). **Interactive Systems. Design, Specification, and Verification**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003. p. 16–33. ISBN 978-3-540-39929-2.

BARZILAY, R.; MCCULLOUGH, D.; RAMBOW, O.; DECRISTOFARO, J.; KORELSKY, T.; LAVOIE, B. A new approach to expert system explanations. In: **Natural Language Generation**. [*S. l.*: *s. n.*], 1998.

BIRAN, O.; COTTON, C. Explanation and justification in machine learning: A survey. In: **IJCAI-17 workshop on explainable AI (XAI)**. [*S. l.*: *s. n.*], 2017. v. 8, n. 1.

BOARD, D. I. **AI Principles**: Recommendations on the ethical use of artificial intelligence by the department of defense. [*S. l.*], 2019. Disponível em: https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF. Acesso em: 01 mai. 2022.

BRANDÃO, R.; CARBONERA, J.; SOUZA, C. de; FERREIRA, J.; GONçALVES, B.; LEITãO, C. **Mediation Challenges and Socio-Technical Gaps for Explainable Deep Learning Applications**. 2019.

BRASIL. **Lei N.º 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados (LGPD)**. 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm. Acesso em: 4 nov. 2019.

BRENNEN, A. What do people really want when they say they want "explainable ai?"we asked 60 stakeholders. In: **Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2020. (CHI EA '20), p. 1–7. ISBN 9781450368193. Disponível em: https://doi.org/10.1145/3334480.3383047. Acesso em: 22 jan. 2021.

BURLE, C.; CORTIZ, D. Mapping principles of artificial intelligence. 11 2019.

CARBONERA, J.; GONÇALVES, B.; SOUZA, C. de. O problema da explicação em inteligência artificial: considerações a partir da semiótica. **TECCOGS: Revista Digital de Tecnologias Cognitivas**, n. 17, 2018.

CARVALHO, N. de O.; SAMPAIO, A. L.; MONTEIRO, I. T. Evaluation of facebook advertising recommendations explanations with the perspective of semiotic engineering. In: **Proceedings of the 19th Brazilian Symposium on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 2020. (IHC '20). ISBN 9781450381727. Disponível em: https://doi.org/10.1145/3424953.3426632. Acesso em: 22 jan. 2021.

COMMISSION, E. **Ethics Guidelines for Trustworthy AI**. [*S. l.*]: Scholar Commons, 2020. Document prepared by the European Union High-Level Expert Group on Artificial Intelligence (AI HLEG). Disponível em: https://scholarcommons.scu.edu/poli_laws_regs/3. Acesso em: 22 abr. 2022.

DARPA. **Explainable Aritificial Intelligence (XAI) Program (DARPA-BAA-16-53)**. 2016. http://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf. Acesso em: 10 mai. 2019.

DASTIN, J. Amazon scraps secret ai recruiting tool that showed bias against women. In: **Ethics of Data and Analytics**. [*S. l.*]: Auerbach Publications, 2018. p. 296–299.

de Souza, C. S.; BARBOSA, S. D. J.; PRATES, R. O. A semiotic engineering approach to user interface design. **Knowledge-Based Systems**, v. 14, n. 8, p. 461 – 465, 2001. ISSN 0950-7051. Semiotic Approaches to User Interface Design. Disponível em: http://www.sciencedirect.com/science/article/pii/S0950705101001368. Acesso em: 22 jan. 2021.

DRESSEL, J.; FARID, H. The accuracy, fairness, and limits of predicting recidivism. **Science Advances**, American Association for the Advancement of Science, [*S. l.*], v. 4, n. 1, 2018. Disponível em: https://advances.sciencemag.org/content/4/1/eaao5580. Acesso em: 22 jan. 2021.

DUDLEY, J. J.; KRISTENSSON, P. O. A review of user interface design for interactive machine learning. **ACM Transactions on Interactive Intelligent Systems (TiiS)**, ACM New York, NY, USA, v. 8, n. 2, p. 1–37, 2018.

EIBAND, M.; SCHNEIDER, H.; BILANDZIC, M.; FAZEKAS-CON, J.; HAUG, M.; HUSSMANN, H. Bringing transparency design into practice. In: **23rd International Conference on Intelligent User Interfaces**. New York, NY, USA: Association for Computing Machinery, 2018. (IUI '18), p. 211–223. ISBN 9781450349451. Disponível em: https://doi.org/10.1145/3172944.3172961. Acesso em: 22 jan. 2021.

EIBAND, M.; SCHNEIDER, H.; BILANDZIC, M.; FAZEKAS-CON, J.; HAUG, M.; HUSSMANN, H. Bringing transparency design into practice. In: **23rd international conference on intelligent user interfaces**. [*S. l.*: *s. n.*], 2018. p. 211–223.

ESCOVEDO, T.; KOSHIYAMA, A. **Introdução a Data Science**: Algoritmos de machine learning e métodos de análise. [*S. l.*]: Casa do Código, 2020.

ETHICS, A. C. on P. **ACM Code of Ethics and Professional Conduct**. 2018. Disponível em: https://www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf. Acesso em: 01 mai. 2022.

EUROPÉIA, U. **Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho de 27 de abril de 2016**. Disponível em: https://eur-lex.europa.eu/legalcontent/PT/TXT/PDF/?uri=CELEX:32016R0679&from=PT. Acesso em: 4 nov. 2019.

FERREIRA, J. J.; MONTEIRO, M. Designer-user communication for xai: An epistemological approach to discuss xai design. **arXiv preprint arXiv:2105.07804**, 2021.

FJELD, J.; ACHTEN, N.; HILLIGOSS, H.; NAGY, A.; SRIKUMAR, M. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. **Berkman Klein Center Research Publication**, n. 2020-1, 2020.

FLORES, A. W.; BECHTEL, K.; LOWENKAMP, C. T. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. **Fed. Probation**, HeinOnline, [*S. l.*], v. 80, p. 38, 2016.

GEBRU, T.; MORGENSTERN, J.; VECCHIONE, B.; VAUGHAN, J. W.; WALLACH, H.; III, H. D.; CRAWFORD, K. Datasheets for datasets. **arXiv preprint arXiv:1803.09010**, 2018.

GÉRON, A. **Mãos à Obra**: Aprendizado de máquina com scikit-learn & tensorflow. [*S. l.*]: Alta Books, 2019.

GOOGLE. **AI at Google: our principles**. 2018. Disponível em: https://www.blog.google/technology/ai/ai-principles/. Acesso em: 01 mai. 2022.

High-Level Expert Group on AI. **Ethics guidelines for trustworthy AI**. Brussels, 2019. Disponível em: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Acesso em: 22 de abril de 2022.

IBM. **Everyday Ethics for Artificial Intelligence**. 2019. Disponível em: https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf. Acesso em: 01 mai. 2022.

INSTITUTE, F. of L. **Asilomar AI Principles**. 2017. Disponível em: https://futureoflife.org/ai-principles/. Acesso em: 01 mai. 2022.

JAKOBSON, R. Linguistics and poetics. In: **Style in language**. [*S. l.*]: MA: MIT Press, 1960. p. 350–377.

LARSON, J.; ANGWIN, J.; MATTU, S.; KIRCHNER, L. **How We Analyzed the COMPAS Recidivism Algorithm**. 2016. Disponível em: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Acesso em: 22 abr. 2022.

LAZAR, J.; FENG, J. H.; HOCHHEISER, H. **Research methods in human-computer interaction**. [*S. l.*]: Morgan Kaufmann, 2017.

LEITE, J. C. Modelos e formalismos para a engenharia semiótica de interfaces de usuário. **Rio de Janeiro**, [*S. n.*], [*S. l.*], 1998.

LIPTON, Z. C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. **Queue**, ACM New York, NY, USA, v. 16, n. 3, p. 31–57, 2018.

LOPES, B. G.; SOARES, L. S.; PRATES, R. O.; GONÇALVES, M. A. Analysis of the user experience with a multiperspective tool for explainable machine learning in light of interactive principles. In: **Proceedings of the XX Brazilian Symposium on Human Factors in Computing Systems**. [*S. l.*: *s. n.*], 2021. p. 1–11.

MICROSOFT. **Microsoft AI Principles**. 2019. Disponível em: https://www.microsoft.com/en-us/ai/our-approach-to-ai. Acesso em: 01 mai. 2022.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. **Artificial Intelligence**, v. 267, p. 1 – 38, 2019. ISSN 0004-3702. Disponível em: http://www.sciencedirect.com/science/article/pii/S0004370218305988. Acesso em: 22 jan. 2021.

MILLER, T.; HOWE, P.; SONENBERG, L. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. **arXiv preprint arXiv:1712.00547**, 2017.

MITCHELL, M.; WU, S.; ZALDIVAR, A.; BARNES, P.; VASSERMAN, L.; HUTCHINSON, B.; SPITZER, E.; RAJI, I. D.; GEBRU, T. Model cards for model reporting. In: . New York, NY, USA: Association for Computing Machinery, 2019. (FAT* '19). ISBN 9781450361255. Disponível em: https://doi.org/10.1145/3287560.3287596. Acesso em: 22 jan. 2021.

MOHSENI, S. Toward design and evaluation framework for interpretable machine learning systems. In: . New York, NY, USA: Association for Computing Machinery, 2019. (AIES '19), p. 553–554. ISBN 9781450363242. Disponível em: https://doi.org/10.1145/3306618.3314322. Acesso em: 22 jan. 2021.

MOLNAR, C. **Interpretable Machine Learning**: A guide for making black box models explainable. Germany, Munich: [*S. n.*], 2020.

MONTEIRO, R. L. Existe um direito à explicação na lei geral de proteção de dados do brasil. **Artigo estratégico**, v. 39, p. 1–14, 2018.

MOORE, J. D.; SWARTOUT, W. R. **Explanation in expert systemss: A survey**. [*S. l.*], 1988.

MUELLER, S. T.; HOFFMAN, R. R.; CLANCEY, W.; EMREY, A.; KLEIN, G. Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. **arXiv preprint arXiv:1902.01876**, 2019.

MUELLER, S. T.; VEINOTT, E. S.; HOFFMAN, R. R.; KLEIN, G.; ALAM, L.; MAMUN, T.; CLANCEY, W. J. **Principles of Explanation in Human-AI Systems**. arXiv, 2021. Disponível em: https://arxiv.org/abs/2102.04972. Acesso em: 22 jan. 2022.

MUNRO, R. **Human-in-the-loop machine learning**. [*S. l.*]: Manning Publications, 2019.

OECD. **OECD Principles on AI**. 2019. Disponível em: https://www.oecd.org/going-digital/ai/principles/. Acesso em: 01 mai. 2022.

OLIVEIRA, E. R. de; PRATES, R. O. Intermediated semiotic inspection method. In: **Proceedings of the 17th Brazilian Symposium on Human Factors in Computing Systems**. Brasil: Association for Computing Machinery, 2018. p. 1–10.

O'NEIL, C. **Weapons of math destruction**: How big data increases inequality and threatens democracy. First edition. New York: Crown, 2016. ISBN 978-0-451-49733-8.

PEIRCE, C. S. Logic as semiotic: The theory of signs. **Philosophical writings of Peirce**, p. 100, 1902.

PEQUIM, A. de Inteligência Artificial de. **Beijing AI Principles**. 2019. Disponível em: https://www.baai.ac.cn/news/beijing-ai-principles-en.html. Acesso em: 22 de abril de 2022.

PRATES, R.; BARBOSA, S. Capítulo 6 introdução à teoria e prática da interação humano computador fundamentada na engenharia semiótica. 10 2020.

PRATES, R. O.; BARBOSA, S. D. J. Introdução à teoria e prática da interação humano computador fundamentada na engenharia semiótica. **Atualizações em informática**, p. 263–326, 2007.

PRATES, R. O.; SILVA, R. F. da. Avaliação do uso da manas como ferramenta epistêmica no projeto de sistemas colaborativos. In: **Proceedings of the IX Symposium on Human Factors in Computing Systems**. Porto Alegre, BRA: Brazilian Computer Society, 2010. (IHC '10), p. 21–30.

SAMPAIO, A. L. **Um Modelo para Descrever e Negociar Modificaçoes em Sistemas Web**. Tese (Doutorado) – PUC–Rio, 2010.

SANTAELLA, L.; NOTH, W. **Introdução à Semiótica**. [*S. l.*]: Paulus Editora, 2021. (Introduções). ISBN 9786555623604. Disponível em: https://books.google.com.br/books?id=3D9OEAAAQBAJ.

SANTEE, N. R.; TEMER, A. C. R. P. A linguística de roman jakobson: Contribuições para o estudo da comunicação. **Revista de Ensino, Educação e Ciências Humanas**, v. 12, n. 1, 2011.

SCHÖN, D. The reflective practitioner. **New York**, v. 1083, 1938.

SCHÖN, D.; BENNETT, J. Reflective conversation with materials. In: ACM. **Bringing design to software**. [*S. l.*], 1996. p. 171–189.

SILVEIRA, M. S.; BARBOSA, S. D.; SOUZA, C. S. de. Model-based design of online help systems. In: JACOB, R. J.; LIMBOURG, Q.; VANDERDONCKT, J. (Ed.). **Computer-Aided Design of User Interfaces IV**. Dordrecht: Springer Netherlands, 2005. p. 29–42.

SOUZA, C. S. D. A pragmatic turn in computer science. **Interactions**, Association for Computing Machinery, New York, NY, USA, v. 25, n. 3, p. 20–21, abr. 2018. ISSN 1072-5520. Disponível em: https://doi.org/10.1145/3200147. Acesso em: 22 jan. 2021.

SOUZA, C. S. D.; LEITÃO, C. F. Semiotic engineering methods for scientific research in hci. **Synthesis Lectures on Human-Centered Informatics**, Morgan & Claypool Publishers, v. 2, n. 1, p. 1–122, 2009.

SOUZA, C. S. D.; LEITÃO, C. F.; PRATES, R. O.; SILVA, E. J. D. The semiotic inspection method. In: **Proceedings of VII Brazilian symposium on Human factors in computing systems**. [*S. l.*: *s. n.*], 2006. p. 148–157.

SOUZA, C. S. D.; NARDI, B. A.; KAPTELININ, V.; FOOT, K. A. **The semiotic engineering of human-computer interaction**. [*S. l.*]: MIT press, 2005.

SOUZA, C. S. de. A pragmatic turn in computer science. **Interactions**, ACM New York, NY, USA, v. 25, n. 3, p. 20–21, 2018.

SWARTOUT, W.; PARIS, C.; MOORE, J. Explanations in knowledge systems: Design for explainable expert systems. **IEEE Expert**, IEEE, v. 6, n. 3, p. 58–64, 1991.

SWARTOUT, W. R. Xplain: A system for creating and explaining expert consulting programs. **Artificial intelligence**, Elsevier, v. 21, n. 3, p. 285–325, 1983.

SWARTOUT, W. R.; MOORE, J. D. Explanation in second generation expert systems. In: **Second generation expert systems**. [*S. l.*]: Springer, 1993. p. 543–585.

TINTAREV, N.; MASTHOFF, J. Explaining recommendations: Design and evaluation. In: RICCI, F.; ROKACH, L.; SHAPIRA, B. (Ed.). **Recommender Systems Handbook**. Springer, 2015. p. 353–382. Disponível em: https://doi.org/10.1007/978-1-4899-7637-6_10. Acesso em: 22 jan. 2021.

TOREINI, E.; AITKEN, M.; COOPAMOOTOO, K.; ELLIOTT, K.; ZELAYA, C. G.; MOORSEL, A. V. The relationship between trust in ai and trustworthy machine learning technologies. In: **Proceedings of the 2020 conference on fairness, accountability, and transparency**. [*S. l.*: *s. n.*], 2020. p. 272–283.

TOREINI, E.; AITKEN, M.; COOPAMOOTOO, K. P.; ELLIOTT, K.; ZELAYA, V. G.; MISSIER, P.; NG, M.; MOORSEL, A. van. Technologies for trustworthy machine learning: A survey in a socio-technical context. **arXiv preprint arXiv:2007.08911**, 2020.

UNION, U. G. Top 10 principles for ethical artificial intelligence. **Nyon, Switzerland**, 2017.

WACHTER, S.; MITTELSTADT, B.; FLORIDI, L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. **International Data Privacy Law**, Oxford University Press, v. 7, n. 2, p. 76–99, 2017.

WEITZ, K.; SCHILLER, D.; SCHLAGOWSKI, R.; HUBER, T.; ANDRE, E. "let me explain!": exploring the potential of virtual agents in explainable ai interaction design. **Journal on Multimodal User Interfaces**, v. 15, 07 2020.

ZHOU, T.; SHENG, H.; HOWLEY, I. Assessing post-hoc explainability of the bkt algorithm. In: **Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society**. [*S. l.*: *s. n.*], 2020. p. 407–413.