UNIVERSIDADE FEDERAL FLUMINENSE
INSTITUTO DE COMPUTAÇÃO
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

[REDACTED] MONTEIRO FERNANDES CORRÊA

UMA AVALIAÇÃO DO TWITTER BASEADA EM PRIVACIDADE, SEGURANÇA, EXPLICABILIDADE, NÃO DISCRIMINAÇÃO E PROMOÇÃO DE VALORES HUMANOS

Niterói
2021

[REDACTED] MONTEIRO FERNANDES CORRÊA

UMA AVALIAÇÃO DO TWITTER BASEADA EM PRIVACIDADE, SEGURANÇA, EXPLICABILIDADE, NÃO DISCRIMINAÇÃO E PROMOÇÃO DE VALORES HUMANOS

Trabalho de conclusão de curso apresentado ao curso de Bacharelado em Ciência da Computação, como requisito parcial para conclusão do curso.

Orientadora
Prof.ª Dr.ª Luciana Cardoso de Castro Salgado

Niterói
2021

AGRADECIMENTOS

Gostaria de agradecer primeiramente aos meus pais, por terem me dado a vida, lutado pela minha saúde e me dado todo o apoio financeiro para que eu pudesse ter não apenas uma educação de qualidade, mas também estar em ambientes que me possibilitem atingir a liberdade.

Segundamente, agradeço ao meu irmão por inúmeras vezes ser um exemplo de ser humano para que eu pudesse ver que o céu não é o limite para o potencial de um indivíduo, além de me proteger a apoiar em situações ameaçadoras e debilitantes.

Terceiramente, gostaria de agradecer aos meus amigos que me apoiaram tanto quando eu quis diversas vezes desistir de minha graduação. Fosse com apoio emocional, psicológico, sugestões ou soluções, eles sempre estavam lá para me motivar.

Em quarto lugar, mas não menos importante que os outros, agradeço muitíssimo a minha orientadora Luciana Salgado por ter acreditado em mim e me guiado por toda a escuridão que foi o início deste trabalho, até que ganhasse confiança para produzir sem precisar consultá-la com frequência.

Finalmente, gostaria de agradecer a mim mesma, por não ter desistido de ser feliz e alcançar meus objetivos, ainda que não acreditasse nessas possibilidades.

“É difícil entender o universo se você estuda
apenas um planeta.”
(Miyamoto Musashi)

RESUMO

O aumento do uso das Tecnologias de Informação e Comunicação no século 21, principalmente a Internet, faz surgir e crescer o fenômeno do *Big Data* e das redes sociais. Este fenômeno está diretamente relacionado à participação ativa das redes sociais na geração de enormes quantidades de dados ao redor do mundo. Para que todo esse volume de dados seja útil e crie benefícios para a sociedade e organizações, é preciso realizar processos e análises de dados que fazem uso de Inteligência Artificial, Inteligência de Negócios e outras práticas. Dessa forma, as organizações que gerenciam as redes sociais são capazes, por exemplo, de melhorar o serviço oferecido para seus usuários. Entretanto, esses processos e análises podem gerar experiências de uso negativas para as milhões de pessoas usuárias das redes sociais, como por exemplo, o contato com discursos de ódio na rede social, a comercialização de dados não-consentida pelos usuários e a disseminação em massa de notícias falsas dentro da rede social. Portanto, o monitoramento constante de possíveis problemas ainda não percebidos pelas pessoas usuárias, projetistas e desenvolvedoras da rede social é necessário, a fim de corrigi-los o quanto antes e evitar uma experiência de uso negativa. Este estudo investigou a rede social Twitter, avaliando a comunicabilidade do Twitter sobre os seguintes princípios éticos para Inteligência Artificial: privacidade, segurança, explicabilidade, transparência, não-discriminação e promoção de valores humanos, além de como os usuários percebem (ou não) esses princípios éticos e o entendimento deles sobre ética. Esses objetivos foram atingidos por meio de dois estudos: i) exploração da comunicabilidade do Twitter com a aplicação do Método de Inspeção Semiótica guiada pelos princípios éticos para Inteligência Artificial; ii) pesquisa (*survey*) com usuários do Twitter para identificar suas percepções sobre problemas éticos que eles encontram durante o uso desta rede social. Os principais resultados indicam potenciais violações éticas causadas pelo design do Twitter, além de revelar o entendimento sobre o que é ética para as pessoas usuárias do Twitter e a influência de Inteligência Artificial no processo de violação de princípios éticos.

Palavras-chave: Engenharia Semiótica. Ética. Inteligência Artificial. Interação Humano-Computador.

ABSTRACT

The increased use of Information and Communication Technologies in the 21st century, mainly the Internet, creates and grows the *Big Data* phenomenon and social networks. This phenomenon is directly related to the active participation of social networks in creating vast quantities of data worldwide. For this data to be helpful and create benefits and solve societies' and organizations' problems, processes and data analysis need to be done using Artificial Intelligence, Business Intelligence, and other practices. Therefore, the organizations that manage their social networks can, for example, improve the services offered to their users. However, these practices can develop negative user experiences to millions of social network users, such as the contact with hate discourses in social networks, commercialization of user data without consent from the users, and massive spread of fake news inside the social network. Thus, monitoring potential problems not yet noticed by the users, designers, or developers of the social network is necessary to fix them as soon as possible to avoid a negative user experience. This study investigated the social network Twitter, evaluating the communicability of Twitter on the following ethical principles for Artificial Intelligence: privacy, security, explanability, transparency, non-discrimination and promotion of human values, in addition to how users perceive (or not) these ethical principles and their understanding of ethics. This objective was achieved through two studies: i) exploring some already known Twitter problems with the Semiotic Inspection Method oriented by ethical principles for Artificial Intelligence; ii) a survey with Twitter users to identify which ethical issues they encounter during the use of the social network. The main results indicate many ethical problems that Twitter's design does not prevent, besides revealing the Twitter users' understanding of ethics and the influence of Artificial Intelligence in the infringement of the ethical principles.

Keywords: Semiotic Engineering. Ethics. Artificial Intelligence. Human-Computer Interaction.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS

GDPR	Global Data Protection Regulation
IA	Inteligência Artificial
IHC	Interação Humano-Computador
LGPD	Lei Geral de Proteção de Dados
MIS	Método de Inspeção Semiótica
TIC	Tecnologia da Informação e Comunicação

SUMÁRIO

1. INTRODUÇÃO

Nas primeiras décadas do século 21, pode-se notar um crescimento no campo de *Big Data* (SAGIROGLU & SINANC, 2013). Isso ocorre devido ao aumento do uso de dispositivos computacionais como computadores e *smartphones*, e a popularização da internet e mídias sociais. Tais tecnologias geram grandes quantidades de dados pessoais que são utilizados por indivíduos ou organizações para desenvolverem soluções para problemas em diversos domínios, a exemplo de Ciência da Computação, Medicina, Biologia e Empreendedorismo (RODRÍGUEZ-MAZAHUA et al., 2015).

Big Data é descrita por Lomotewy e Deters (2014) a partir do modelo 5V, o qual define os conceitos de Volume (grande quantidade de dados gerados), Velocidade (a informação deve ser obtida em tempo real, idealmente), Variedade (os dados podem ser estruturados, semiestruturados e não-estruturados), Veracidade (os dados precisam ser verificados quanto a sua veracidade) e Valor (a coleta e análise de dados devem contribuir relevantemente para o problema a ser solucionado).

O objetivo do *Big Data* é criar sistemas de processamento, armazenamento e análise de dados para eficientemente lidar com esses dados e analisá-los para extrair significados relevantes para apoiar tomadas de decisão de organizações como empresas e governos. Além disso, métodos de análise de *Big Data* têm sido utilizados junto de métodos de IA para analisar conjuntos grandes de dados (RODRÍGUEZ-MAZAHUA et al., 2015).

As redes sociais são tecnologias notáveis na Ascensão do *Big Data* no período citado, pois recebem um grande fluxo de acessos simultâneos, em sua maioria de pessoas, o que gera uma quantidade enorme de dados que dizem respeito a estes indivíduos. Esses dados, por sua vez, são processados, organizados e armazenados pela rede social e podem ser disponibilizados para entidades externas, dependendo das políticas de compartilhamentos de dado da rede social.

De qualquer forma, com a disponibilidade de informações sobre indivíduos ou grupos, é possível extrair informações sobre sociedades, culturas e indivíduos que vão de encontro ao interesse de diversos âmbitos como política, saúde e economia e instituições que atuam neles.

Estima-se que as mídias sociais possuam 4.33 bilhões de usuários, o que representa em torno de 55% da população global, que ao todo gasta 10 bilhões de horas por dia em redes

sociais (DATAREPORTAL, 2021). O *Instagram*©[1], por exemplo, gera 100 milhões de curtidas por hora (CAREY-SIMOS, 2021), ou seja, 100 milhões unidades de informação sobre interesse das pessoas que curtiram uma foto. O Twitter©[2], por sua vez, possui 187 milhões de acessos de usuários por dia (DEAN, 2021). Com a formação em alta velocidade de grande e variado volume de dados sobre indivíduos ou grupos, é possível extrair informações sobre sociedades, culturas e indivíduos que são de interesse de diversos segmentos como política, saúde, economia e instituições que atuam neles.

Outro aspecto importante é que os dados produzidos e capturados se transformam em bens que são negociados entre diversas partes. Empresas e outras instituições costumam armazenar os dados de seus usuários e em determinados momentos, vendem-nos para outras empresas ou instituições, que podem utilizar essas informações para melhorar seus processos ou atividades e obter algum valor a partir disso. Uma prática comum é o uso dos dados pessoais para alavancar estratégias de publicidade digital (E-COMMERCE BRASIL, 2019). Em 2019, as empresas estadunidenses gastaram 33,73 bilhões de dólares em anúncios dentro de redes sociais, configurando um aumento de 18% em relação à 2018 (E-COMMERCE BRASIL, 2019). Em 2018, no Brasil, 50,9% dos lojistas utilizavam anúncios no Facebook©[3] e 46,7% no Instagram, notando um aumento de 47,9% nas vendas após o uso dessa estratégia (E-COMMERCE BRASIL, 2019). Isso mostra que o uso de dados pessoais para publicidade tem crescido e é cada vez mais relevante para empresas.

As redes sociais costumam utilizar IA para, dentre outras tarefas, recomendar conteúdo para seus usuários consumirem. Através da determinação dos interesses de um usuário, as redes são capazes de mostrar conteúdo relacionado a eles. Uma das formas de determinar o interesse de um usuário é através da sua interação com um determinado tópico ou assunto. A partir disso, os algoritmos da rede buscam conteúdo relacionado ao tópico ou assunto para mostrar ao usuário (WARREN, 2019).

1.1.Ética e *Big Data*

Embora o uso de IA e *Big Data* tenha seus benefícios, existem potenciais problemas éticos relacionados a essa combinação de tecnologias, tanto no processo de coleta e armazenamento dos dados das atividades dos usuários no sistema quanto nos algoritmos de IA que os utilizam.

[1] instagram.com (última visita em 24/04/2021)

[2] twitter.com (última visita em 24/04/2021)

[3] facebook.com (última visita em 24/04/2021)

A exemplo, a Microsoft criou um perfil no Twitter para o seu robô de conversação, Tay. A princípios, a Tay interagia com os usuários e aprendia sobre assuntos a partir dessas interações. Após algum tempo interagindo com outros usuários no Twitter, Tay começou a exibir discursos de ódio e apologia aos movimentos fascistas, ocasião em que a Microsoft ® começou a apagar as postagens do perfil do robô (VINCENT, 2016).

Outro exemplo é a falha dos mecanismos automáticos para detectar discursos de ódio na plataforma. Um estudo sobre a eficácia dos métodos de detecção de discursos de ódio direcionados a mulheres descobriu que, embora existam ferramentas para descobrir e punir tais falas, ele ainda apresenta falhas ao detectar discursos que continham palavras-chave ainda não utilizadas, ou seja, que foram criadas pelos ofensores para contornar esses mecanismos (PAVLIUC, 2021). Além disso, no Twitter, existem dezenas de milhares de discursos de ódio, mostrando que, de fato, esses mecanismos não são confiáveis, como mostrado por ElSherief et al. (2018).

Há também a questão da privacidade e segurança dos usuários. A partir da interação com as redes sociais, dados são produzidos e devem ser protegidos para que as pessoas não sofram consequências indesejadas. Uma das iniciativas que visam amenizar esse problema são as Leis Gerais de Proteção de Dados (LGPD), no Brasil, e a Global Data Protection Regulation (GDPR), na Comunidade Europeia (LGPD BRASIL, 2021; NADER, 2020), que definem medidas mínimas a serem implementadas por instituições em seus sistemas para que seus usuários tenham controle (agência) sobre seus dados e estejam protegidos em caso de alguma falha de segurança por parte da instituição que os detém.

Entretanto, são inúmeras as ocorrências que denotam a falta de segurança ou privacidade nas redes sociais (OHLHORST, 2021). Em 2020, por exemplo, um invasor obteve as credenciais de um funcionário do Twitter com permissões elevadas no sistema. A partir dessas credenciais, o atacante conseguiu publicar postagens em diversos perfis famosos, como o do ex-presidente americano Barack Obama e do empresário Bill Gates, além de baixar os dados das contas (REED, 2020). Com isso, é possível, para os invasores, obterem valor a partir dos dados obtidos, podendo até vendê-los para outras instituições, já que as informações são de perfis de extrema relevância atualmente.

Os problemas mencionados estão diretamente relacionados aos princípios éticos para IA de privacidade, segurança, transparência, não-discriminação e promoção de valores humanos definidos no trabalho de Fjeld et al. (2020). Fundamentalmente, os princípios éticos tentam proteger os usuários de um sistema que utiliza IA de potenciais danos, ao passo que

alinham o desenvolvimento de um modelo de IA com valores éticos e morais de uma sociedade.

1.2.Objetivos

Este trabalho tem como objetivo avaliar a comunicabilidade do Twitter sobre os princípios éticos para IA de privacidade e segurança, transparência e explicabilidade, não-discriminação e promoção de valores humanos, implementados na tecnologia, e como esses usuários percebem (ou não) esses cuidados éticos.

Comunicabilidade é "a propriedade do software que eficientemente e efetivamente transmite aos usuários o design subjacente e princípios de interação" (PRATES, DE SOUZA & BARBOSA, 2000).

Mais especificamente, os objetivos são:

- Revelar potenciais problemas na comunicabilidade da ética pelo Twitter, em situações de uso da aplicação.
- Entender se as pessoas usuárias do Twitter, de diferentes perfis, enfrentam ou percebem problemas éticos durante o uso cotidiano.
- Entender a percepção das pessoas usuárias do Twitter sobre Ética.

1.3.Metodologia

O trabalho foi desenvolvido utilizando uma metodologia descritiva qualitativa, a partir de inspeções informais, da aplicação do Método de Inspeção Semiótica (MIS) (DE SOUZA et al., 2006), e de uma pesquisa (*survey*) com pessoas usuárias do Twitter. A seguir está a organização dos procedimentos empregados:

- Etapa 1: Inspeção preliminar na interface para identificar possíveis focos de discrepância e escolha dos princípios éticos para IA para realização de análises mais profundas;
- Etapa 2: Aplicação do MIS, buscando por potenciais rupturas de comunicabilidade pelas estratégias comunicativas relacionados aos princípios éticos para IA;
- Etapa 3: *Survey* com pessoas usuárias, buscando complementar e contrastar os resultados obtidos nas etapas anteriores; e

- Etapa 4: Comparação dos resultados do *survey* com os resultados das inspeções.

1.4. Organização do Trabalho

No Capítulo 2, é descrita a fundamentação teórica do trabalho, são apresentados os princípios éticos para IA escolhidos para avaliar o Twitter, assim como uma discussão de Ética e IA.

No Capítulo 3, são discutidos trabalhos relacionados ao tema da avaliação e design ético para sistemas inteligentes.

No Capítulo 4, apresenta-se os resultados da inspeção preliminar e do MIS; e também da entrevista realizada com os usuários para validar e estender os resultados anteriormente.

No Capítulo 5, conclui-se o trabalho e discutem-se as limitações dele e possíveis trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo discute os conceitos de Ética e IA, de Fjeld et al. (2020) e também de Cutler, Pribić e Humphrey (2019), assim como os princípios éticos para IA de Privacidade e Segurança, Transparência e Explicabilidade, Não-discriminação e Promoção de Valores Humanos presentes no trabalho de Fjeld et al. (2020).

Os princípios de Privacidade e Segurança, assim como os de Transparência e Explicabilidade, são analisados juntos. Isso ocorre pois, na seção 2.2, entende-se que a redução (ou aumento) de privacidade também reduz (ou aumenta) a segurança de um indivíduo. E na seção 2.3, entende-se que a redução (ou aumento) da explicabilidade reduz (ou aumenta) a transparência de um algoritmo ou de uma interface.

2.1.Ética e Inteligência Artificial

Fjeld et al. (2020) afirmam que não há um conceito consolidado de IA, mas cita algumas definições encontradas na literatura. Resumidamente, de acordo com os autores, é possível entender IA como sistemas que são capazes de tomar ações de forma autônoma para atingir um objetivo, utilizando aprendizado e raciocínio. Neste trabalho, será utilizado esse conceito quando forem mencionados algoritmos de IA, aprendizado de máquina ou apenas IA.

Martinez (2019) concorda com a diversidade de definições do termo, mas comenta que a maioria das pesquisas em IA busca determinar e imitar a inteligência humana. A definição também depende de quatro fatores: pensar e agir como um humano; e pensar e agir racionalmente. O autor alega que os quatro atributos deveriam ser legalmente adotados. Não obstante, os diversos fatores e definições convergem para a capacidade de ser autônomo.

Um subcampo de IA é o Aprendizado de Máquina. Jordan e Mitchell (2015) descrevem Aprendizado de Máquina como a disciplina que tenta “construir tecnologias e sistemas que evoluem e/ou aprendem automaticamente por meio da experiência em agir dentro de um ambiente, são compostos por modelos estatísticos e utilizam dados para o seu aprendizado”. Esse campo impacta diversas outras esferas de estudo, como biologia, economia, ciência da computação, ciências sociais e, como visto neste trabalho, ética. Os autores questionam quais comportamentos e decisões de uma IA devem ser promovidos e quais devem ser rejeitados pela sociedade, gerando um regulamento sobre IA, devido ao crescimento na adoção dessa tecnologia.

Nesse ponto, Floridi et al. (2018) consideram que a IA “não deve ser regulada apenas após estar em uso ou após amadurecer, já que possui um impacto social muito grande”. Tais autores sugerem o questionamento de quem a IA irá afetar, além de como, onde e quando, tanto negativamente quanto positivamente. Concluem, então, que IA deve ser responsável por suas ações e não causar danos às pessoas usuárias. Além disso, o fato de ser vista e/ou disponibilizada como uma “caixa preta” dificulta que as pessoas entendam e confiem nela.

Floridi e colegas salientam que “a adoção de um framework ético para o desenvolvimento de uma IA possibilita o aproveitamento dos benefícios positivos dela para a sociedade, e mitigação ou prevenção das consequências negativas e socialmente rejeitadas”.

Cutler, Pribić e Humphrey (2019) também argumentam pela inclusão da ética desde o início do processo de desenvolvimento da IA. Neste caso, a IA deve ser desenvolvida de acordo com os princípios éticos e valores sociais afetados por sua implantação e uso. Para tanto, “os designers e desenvolvedores envolvidos em sua criação devem compreender os aspectos éticos e o estado da arte da literatura sobre ética durante todo o processo de desenvolvimento, incluindo a manutenção, para que os humanos confiem nas máquinas”.

Os autores consideram cinco princípios que devem guiar o desenvolvimento ético de uma IA: responsabilidade, alinhamento de valores, explicabilidade, justiça ou não-discriminação e direito dos usuários ao controle dos seus dados. Os autores alegam que todos esses pontos têm definições consolidadas ou bem discutidas na literatura, que podem servir de referência para a construção da IA.

No geral, Cutler, Pribić e Humphrey (2019) entendem que ética são as definições de certo e errado que os humanos devem seguir, e são descritas em termos de direitos, deveres, benefícios para a sociedade, não-discriminação e virtudes específicas. Este trabalho irá, então, utilizar esse conceito de ética para discutir e avaliar o objeto de estudo.

Nas seções e capítulos seguintes, serão apresentados problemas de ética em IA relacionados aos princípios éticos para IA de privacidade, segurança, transparência, explicabilidade, não-discriminação e promoção de valores humanos que constam no trabalho de Fjeld et al. (2020), utilizados para fundamentar os estudos realizados no capítulo 4.

2.2.Privacidade e Segurança

Westin (1967, apud VAN DEN HOVEN et al., 2014) define privacidade como “o direito que indivíduos têm de determinar até que ponto os outros podem saber informações sobre eles”. Todavia, “eles podem não saber que tem direito à certos aspectos de privacidade”,

e além disso podem ser “coagidos a abrir mão de alguns direitos, mesmo em casos onde exista proteção legal para tal” (LAUFER & WOLFE, 1977). Nota-se que a falta de aviso de coleta de dados a um indivíduo pode ser uma ferramenta para obter informações de um indivíduo que até então eram privadas, já que o indivíduo não sabe que seus dados estão sendo coletados.

Sobre privacidade, também é possível afirmar que:

> A privacidade garante que todos possam controlar as informações que dizem respeito a si mesmo, formulando seus próprios conceitos de identidade pessoal, valores, preferências, objetivos, e protegendo suas escolhas do controle público, desgraça social ou objetificação do indivíduo”. Assim, para os autores, a privacidade se torna então um componente fundamental para a construção de um indivíduo. (GRITZALIS et al., 2014).

Ademais, a privacidade seria necessária para manter a dignidade e respeito da condição humana. Dignidade é a “habilidade de se expressar livremente dentro de uma sociedade, de reconhecer a personalidade de um indivíduo e não interferir com as escolhas de vida de outro” (GRITZALIS et al., 2014). Novamente, a privacidade é imprescindível para que um indivíduo venha a se desenvolver adequadamente.

O conceito de privacidade evoluiu junto das Tecnologia da Informação e Comunicação (TICs), ou seja, sistemas para armazenar, processar e distribuir informação, sendo difícil separar as duas esferas (VAN DEN HOVEN et al., 2014). Van den Hoven et al. (2008, apud VAN DEN HOVEN et al., 2014) determinam alguns motivos morais para proteção de dados pessoais, que os quais podem ser ligados ou são ligados ou não a um indivíduo (VAN DEN HOVEN et al., 2014):

- Ausência de prevenção de danos, pois o acesso a dados pessoais pode ser utilizado para causar danos ao indivíduo ao qual pertencem;
- Desigualdade informacional, onde indivíduos não possuem meios de negociar contratos sobre o uso de seus dados e checar se as partes estão cumprindo com os termos do contrato;
- Discriminação e injustiça informacional, quando informações obtidas através de um meio (como medicina) pode ser utilizado em outras esferas (como transações comerciais), levando a desvantagens e discriminações para o indivíduo; e
- Invasão da autonomia moral e dignidade humana, decorrente da falta de privacidade, que expõe indivíduos a forças externas que podem influenciar suas escolhas e fazê-los tomarem decisões que outrora não tomariam. E de

> acordo com Bruynseels e Van den Hoven (2015, apud VAN DEN HOVEN et al., 2014), a possibilidade de se determinar um indivíduo a partir de seus dados pode ser considerada uma falta de modéstia moral.

A Internet, uma tecnologia capaz de distribuir informação em larga escala, foi inicialmente projetada na década de 1960 para ser "uma rede de troca de informações científicas entre pessoas que se conheciam na vida real" (MICHENER, 1999; ELLISON, 2007, apud VAN DEN HOVEN et al., 2014). Assumia-se na época que não haveriam danos causados pelo uso de grandes redes sociais, e com isso as preocupações com privacidade e segurança apareceram apenas mais tarde, sendo solucionadas por meio da construção de *add-ons*[4] ao invés de estarem embutidas no design original de seus criadores (VAN DEN HOVEN et al., 2014).

O conceito de privacidade em redes sociais requer maior aprofundamento. As redes sociais requerem soluções para seus próprios desafios morais, não apenas para limitar o acesso e a divulgação dos dados pessoais dos indivíduos, mas também para limitar a influência das redes sobre as pessoas usuárias. Isto pode acontecer, por exemplo, com a exposição do indivíduo a variados tipos de informação e funcionalidades, que podem causar danos à sua privacidade e segurança, à longo prazo (VAN DEN HOVEN et al., 2014).

Para Van den Hoven et al. (2014), "a rede influencia seus usuários a trocar seus dados pessoais por certos benefícios, e eles podem não estar cientes da informação que estão cedendo". Dentro desse contexto, "limitar o acesso às informações pessoais da pessoa usuária por outras pessoas não seria suficiente para protegê-las, necessitando de medidas que intervenham na vontade do usuário em compartilhá-las" (VAN DEN HOVEN et al., 2014), que muitas vezes fazem o impulsivamente.

Smith et al. (2012, p. 2) argumentam que "a quantidade de mídia social submetida à Internet está crescendo rapidamente, ao ponto de que é quase impossível para um indivíduo acompanhar o conteúdo submetido por todos os outros usuários". Dizem ainda que "mesmo que existam controles para gerenciar a privacidade do conteúdo ligado ao próprio indivíduo, faltam medidas para que eles possam tratar das complicações geradas após consumirem o conteúdo dos outros usuários da rede" (SMITH et al., 2012). Assim, percebe-se que é difícil para um usuário decidir qual conteúdo ele deseja ver, o que pode ou não o prejudicar consideravelmente.

[4] *Add-ons* são extensões capazes de prover funcionalidades adicionais a aplicações.

De acordo com Smith et al. (2012), os problemas de privacidade podem ser divididos em duas classes: a primeira, onde um usuário "U" submete alguma informação sem medidas de privacidade e proteção suficientes, causando algum dano à ele e sua privacidade; e a segunda, onde um conteúdo que contenha alguma informação sobre um usuário "U" é submetido por outra pessoa, comprometendo a privacidade e segurança de "U". Ainda, para a segunda classe, é necessário que o conteúdo possa ser ligado de alguma forma ao usuário "U", e que o conteúdo, de fato, cause algum dano perceptível à "U".

Beigi e Liu (2020) também relatam dois tipos de problemas de privacidade, que ocorrem quando dados são publicados por um provedor de serviço externo: a divulgação de identidade, que ocorre quando um indivíduo pode ser relacionado a um dado dentro do conjunto de dados publicado; e a divulgação de atributo, onde é possível inferir alguma informação sobre um indivíduo a partir dos dados publicados.

Um problema notável decorrente da publicação dos dados é a geolocalização de um indivíduo através do conteúdo publicado em redes sociais, onde uma pessoa pode, através da informação, prever o movimento de um indivíduo e identificar pontos de interesse (BEIGI & LIU, 2020).

Outro problema comum é causado pelos sistemas de recomendação, i.e. aplicações capazes de determinar os conteúdos preferidos por um usuário. Embora tais sistemas melhorem o serviço oferecido a um usuário, levantam preocupações relacionadas a privacidade e segurança ao representarem essas preferências. Nesses casos os interesses do usuário são registrados e serão expostos caso não sejam propriamente protegidos (BEIGI & LIU, 2020).

Gritzalis et al. (2014) argumentam que existem métodos para determinar tipos de perfis de usuário através do processamento do conteúdo submetido por eles na rede. De acordo com os autores, os usuários não percebem que seus dados estão sendo utilizados para vários motivos, e ressaltam ainda que outros pesquisadores notam a emergência de perigos sociais através do perfilamento[5] dos dados. Beigi e Liu (2020) também revisam diversos métodos criados para inferir ou obter informações a partir de dados publicados, mostrando que há várias maneiras para que um indivíduo sofra danos a partir da sua exposição.

[5] Perfilamento, nesse contexto, é a determinação ou identificação de um perfil ou indivíduo através de informações relacionadas a esse perfil ou indivíduo.

Para lidar com esses e outros problemas, é necessário que os dados dos usuários sejam sanitizados[6], como, por exemplo, por meio da anonimização, i.e. a remoção ou alteração dos dados, de forma a dificultar a inferência de informações sensitivas por indivíduos interessados ou aproveitadores (BEIGI & LIU, 2020).

2.3.Transparência e Explicabilidade

Para Plaisance (2007), a transparência pode ser considerada uma questão "não relacionada apenas ao que é dito, mas também ao porquê e como algo é dito": De acordo com o autor, a transparência:

> Nos torna seres autônomos e racionais que são capazes de avaliar comportamentos uns dos outros. Um comportamento transparente pode ser definido como a conduta que presume uma comunicação aberta, onde trocas entre diversos participantes devem ser francas, caso eles possuam influência legítima e efetiva sobre as consequências decorrentes da comunicação. Não necessariamente é proibido omitir ou mentir, o que pode acontecer caso sejam necessários para satisfazer necessidades de privacidade. Mesmo assim, o conceito de transparência garante que todos os participantes saibam o que está sendo dito ou feito, de forma que ocorram trocas honestas de informação (PLAISANCE, 2007).

Esse conceito prevê, por exemplo, que um usuário de uma aplicação que utiliza algoritmos de IA saiba o que está acontecendo, o motivo das decisões serem tomadas e como foram tomadas, englobando aspectos de transparência e de explicabilidade. Para isso, os projetistas da aplicação devem comunicar (explicar) ao usuário como funciona o algoritmo ou aplicação de forma honesta. Em suma, projetistas e desenvolvedores devem promover a explicabilidade de suas aplicações para seus usuários, esclarecendo o seu funcionamento. Tais conceitos são, portanto, fundamentais para a investigação conduzida neste estudo.

Há também outras definições de transparência na literatura. Para Turilli e Floridi (2009), "a transparência pode ser entendida de pelo menos duas formas, ambas dissonantes entre si". Em estudos na área de ética da informação, por exemplo, transparência trata da visibilidade da informação, que pode ser revelada através da divulgação, estando mais relacionada ao conceito de privacidade visto neste capítulo. Todavia, em campos como ciência da computação, transparência está mais relacionada com a invisibilidade da informação, em situações como um programa ser abstraído para um usuário utilizá-lo.

[6] Sanitizar dados significa remover ou alterar um dado para que ele não possa ser recuperado futuramente.

Para Turilli e Floridi (2009), nem sempre a visibilidade da informação gera consequências éticas, "com alguns dados sendo neutros ou relacionados ao design da aplicação, fundamentais para uma boa interação entre o usuário e o computador, como alertas ao clicar em um botão ou ao receber e-mails". Mesmo assim, a transparência pode se tornar um fator que contribui ou atrapalha a ética, caso influencie outro princípio ético. Assim, a transparência depende de outro(s) princípio(s) ético(s) para ser considerada um também. Alguns exemplos destes princípios são privacidade, segurança, bem-estar e consentimento informado. Os dois últimos muito relacionados com os conceitos de privacidade e segurança analisados neste estudo.

Adicionalmente, Turilli e Floridi (2009) concluem que há contextos em que é interessante para uma organização "promover a transparência por meio da divulgação de como é feito o gerenciamento de informações e processos computacionais" (como em aplicações que fornecem serviços de segurança e privacidade), ao ponto que fique claro se "essas ações estão alinhadas com os princípios éticos adotados pela organização" (TURILLI & FLORIDI, 2009). Embora a transparência seja um conceito muito útil para lidar com práticas éticas em IA, o fato de ser muito dependente de outros princípios a torna difícil de ser analisada individualmente.

Em outro estudo, mais focado em explicabilidade, Coeckelbergh (2020) discute quem deve ser responsável pelas decisões tomadas por uma IA. O autor diz que entender a tecnologia que está sendo usada é importante para a responsabilidade, no contexto de IA. "Ainda que os usuários e desenvolvedores de uma IA saibam o objetivo alcançado por ela, raramente entendem as consequências inesperadas e morais decorrentes dela" (COECKELBERGH, 2020), como vieses nos resultados ou algoritmos. Além disso, podem não ter certeza das consequências para quem sofreu algum dano devido aos vieses. De forma geral, os usuários e desenvolvedores não têm consciência das consequências de suas ações ao interagir com uma IA, até mesmo os mais experientes no assunto.

Assim, Coeckelbergh (2020) questiona como empoderar usuários e desenvolvedores para estarem cientes desses problemas, notando o empecilho de que "a IA é frequentemente entendida como uma 'caixa preta', opaca para os indivíduos externos". Nesse quesito, encontra-se o princípio de explicabilidade: para que seja possível dar consciência das decisões que estão sendo tomadas pelo algoritmo para os respectivos afetados, deve-se ser capaz de explicá-las transparentemente.

Mesmo que o estudo de Coeckelbergh (2020) tenha uma base filosófica bem consolidada para fundamentar explicabilidade (e em certo ponto, transparência), o autor

relaciona a explicabilidade com responsabilidade. Como no trabalho de Floridi e Turilli (2019), explicabilidade e transparência estão relacionados a outro termo da ética, necessitando que ocorra um evento que engloba esse termo (por exemplo, um carro autônomo irresponsavelmente atropelar um pedestre), para que só então seja possível justificar e analisar o ocorrido através da lente de transparência e explicabilidade.

Dessa forma, é difícil analisar esses princípios caso não seja detectado um problema. Já com o conceito exposto por Plaisance (2007) isso é possível, pois ele apenas prevê que a transparência deve ser garantida, não precisando de justificativas ou problemas adicionais para ter seu cumprimento questionado.

2.4.Não-discriminação

O problema de discriminação está muito relacionado à promoção de valores humanos. Para Enteman (1996), "a discriminação intencionalmente emprega estereótipos e preconceitos para agir ou julgar alguém, embora o ator da discriminação negue isso, hipocritamente". Além disso "ao estereotipar alguém, essa pessoa se torna a representação de um grupo maior a qual foi decidido que ela pertence, tendo sua humanidade e individualidade, portanto, negadas" (ENTEMAN, 1996). Através da discriminação, essa negação é amplificada com base na generalização.

Ainda, o autor discorre sobre estudos de Immanuel Kant, compreendendo que "já que pessoas não são objetos, ou seja, não podem ter um valor atribuído, então todos devem ser tratados como pessoas dignas e não manipulados como objetos" (ENTEMAN, 1996). A discriminação infringe essa afirmação, já que desumaniza a pessoa intencionalmente, obtendo algum valor a partir disso.

Assim, este trabalho considera esse aspecto ao analisar questões relacionadas à discriminação: nenhum indivíduo deve ter sua individualidade negada através da generalização de preconceitos e estereótipos. Esse conceito é genérico o suficiente para analisar tanto algoritmos quanto design de interfaces e comunicabilidade (PRATES, DE SOUZA & BARBOSA, 2000).

Wittkower (2016) discute que "para evitar discriminação, uma perspectiva antidiscriminatória deve ser empregada durante a atividade de design". Ele cita situações hipotéticas como um empregador obrigar que todas as aplicações para vagas de emprego fossem submetidas por um sistema incompatível com leitores de tela, o que pode eliminar candidatos por motivos não relacionados ao emprego em si.

Além disso, discute como determinar a discriminação decorrente do design. Apesar de não existirem formas objetivas de se atingir isso, e considerando que teorias abstratas não consigam estabelecer divisões entre discriminação que gera problemas e exclusão inofensiva, é notável que existe um limite para que uma exclusão inofensiva se torne uma discriminação problemática.

Wittkower (2016) diz que na disciplina de Interação Humano-Computador (IHC) e em outras, o estudo de *affordances* (dicas visuais para auxiliar a interação com uma interface) é bem disseminado, ao contrário das *non-affordances*, que são empecilhos capazes de excluir membros de um grupo a que pertencem. As *non-affordances*, ou *disaffordances*, também reagem e dependem de contexto sociais, culturais, individuais, como por exemplo usar dicas sonoras na interface para usuários surdos, se tornando discriminatórias. O autor mostra que a discriminação ocorre mais facilmente quando interagida por um usuário do que durante o seu design através das *affordances* pobres, apelidadas por ele de "design ruim".

O termo descrito por Wittkower (2016) como *dysaffordances* são aquelas *disaffordances* que requerem que um usuário se identifique de forma errada para poder utilizar uma funcionalidade de um produto ou serviço. O autor cita o exemplo de pessoas de gênero não-binário precisarem escolher entre gênero feminino ou masculino para preencher um formulário ou a necessidade de verificar o nome verdadeiro em redes sociais ao invés do nome social, comum entre pessoas transgêneros. Visto isso, a discriminação ocorre de formas diretas ou indiretas.

Contudo, a diversidade das experiências de usuário dificulta a utilização de metodologias consolidadas para aliviar esses problemas. Wittkower (2016) sugere que sejam feitos acompanhamentos com usuários para que sejam detectados e resolvidos problemas de discriminação, como é feito nesse estudo. Adicionalmente, a diversidade de membros em equipes de desenvolvimento pode ajudar a amenizar esse problema, ainda que não resolva completamente.

Uma estratégia discutida é a utilização de variação fenomenológica, onde a partir de uma relação humano-tecnologia inicial, altera-se sistemática e iterativamente diferentes aspectos do humano ou da tecnologia para capturar possíveis mudanças na relação humano-tecnologia. Mesmo assim, de acordo com o autor, não há substituição para incluir pessoas diversas no processo de design.

A disciplina de IA é muito conhecida pelos problemas com vieses em algoritmos, que tentam encontrar padrões e generalizá-los para a criação de um modelo. Nos estudos de Ferrer

et al. (2020) entende-se que os vieses, embora atuem em processos discriminatórios, não necessariamente geram discriminação e são imprescindíveis em certos casos para encontrar padrões em um conjunto de dados.

Mesmo assim, Ferrer e colegas enfatizam três situações em que podem ocorrer vieses discriminatórios: na construção do modelo ou algoritmo de IA; no treinamento dos algoritmos; e durante o uso dos algoritmos. Determinar se ocorreu ou não discriminação vai depender do contexto em que foi empregada, levando em consideração fatores socioculturais e históricos que levaram à sua construção. Um algoritmo também é capaz de reforçar, através da discriminação, desigualdades socioculturais já existentes no contexto em que foi aplicado, por exemplo.

De qualquer maneira, não existe um método padrão para avaliar eticamente vieses em IA, já que depende dos criadores e desenvolvedores dela, mas Ferrer et al. (2020) notam uma crescente necessidade de se melhorar práticas de transparência e explicabilidade em algoritmos de IA. Essa transparência poderia ajudar o público afetado a formarem opiniões sobre a discriminação feita pelo algoritmo.

Ferrer et al. (2020) também argumentam que medidas legais de proteção contra discriminação nem sempre são efetivas. Embora existam leis que cobrem diversos tipos de discriminação, como no mercado de trabalho, outras áreas possuem nuances que permitem que discriminadores as infrinjam sem consequências, como cobrar preços diferentes por um serviço baseado no salário de pessoas diferentes.

Para eles, as leis, a exemplo da LGPD (LGPD BRASIL, 2021) e a GDPR (NADER, 2020) da União Europeia, também não se mostram preparadas para lidar com desafios técnicos. Algumas leis requerem que as implementações dos algoritmos sigam um modelo transparente, o que nem sempre é possível ou não fornece indicadores de discriminação suficientes. Resumidamente, mesmo que se garanta a explicabilidade, pode não se identificar discriminação ou ter certeza da ocorrência dela.

Como se pôde ver, a discriminação apresenta problemas éticos, políticos e tecnológicos. Algumas leis tentam combater a presença de vieses em algoritmos, mas não obtiveram muito sucesso até o momento.

2.5.Promoção de valores humanos

De acordo com a Declaração Universal de Direitos Humanos[7], "todo indivíduo tem direito à vida, liberdade e segurança, assim como liberdade de opinião sem interrupção em qualquer tipo de mídia" (UNICEF, 2020). A Declaração também discute igualdade perante ao sexo, nacionalidade, idade e outros fatores individuais.

Há mais de duas décadas, Metzl (1996) discorreu em seu trabalho sobre como os direitos humanos são impactados pelo uso da informação. Inicialmente, as TICs facilitavam muito os movimentos que lutam por direitos humanos ao acelerar a disseminação da informação através da Internet. Ela possibilitou também comunicações à distância de maneira mais fácil e imediata que telefones e outras tecnologias anteriores.

Por outro lado, o uso das TICs também tem seus lados negativos. Metzel (1996) diz que, devido à enorme quantidade de dados disponíveis sobre indivíduos na rede, seria possível para governos e organizações impactarem a vida dessas pessoas de maneira negativa. Outro exemplo é a obtenção de dados privados por invasores através de falhas de segurança em protocolos da Internet.

Um ponto muito importante na crítica de Metzl (1996) é que conforme as TICs se desenvolvem e modernizam setores como o do comércio, existe o risco de alguns indivíduos perderem acesso a esses setores por causa de inacessibilidade a tecnologia, configurando uma exclusão e restringindo a liberdade individual.

Percebe-se então, que as TICs estão vinculadas aos direitos humanos de indivíduos. O fato dela impactar negativamente a liberdade e segurança de indivíduos viola os conceitos enunciados anteriormente. Porém, ainda que isso ocorra, a tecnologia é capaz de beneficiar indivíduos e deve ser usada para tanto, evitando prejudicá-los.

Como discutido por Buchanan (2001), o design está essencialmente ligado aos valores e aos direitos humanos. Especificamente, o autor declara que o Design Centrado no Humano (NORMAN & DRAPER, 1986) é fundamentalmente a afirmação da dignidade humana e uma tentativa de promovê-la e fortalecê-la independente de contextos sociais, econômicos, políticos e culturais.

De acordo com Buchanan (2001), através do design, é possível criar artefatos para servir aos seres humanos, cumprindo suas necessidades e facilitando a comunicação entre eles. Esses artefatos abrangem desde sistemas de informação até constituições e a qualidade da comunicação, interações e ambientes é a própria expressão dos valores nacionais e culturais onde ocorrem.

[7] Adotada e proclamada pela Assembleia Geral das Nações Unidas (resolução 217 A III) em 10 de dezembro de 1948.

Utilizando esse raciocínio, entende-se que não apenas a disciplina de Design no geral é impactada, mas que também é possível utilizar práticas de design para melhorar algoritmos de IA, como mostra o trabalho de Auernhammer (2020). Tanto interfaces quanto algoritmos, portanto, podem utilizar métodos de design alinhados com o Design Centrado no Humano para promover valores humanos e igualdade perante diversos contextos no produto resultante.

Sellen et al. (2009), entretanto, entendem que o humano é apenas uma parte do processo de Design, existindo também valores sociais e institucionais, que nem sempre estão alinhados com os valores individuais como privacidade e segurança. Isso ocorre devido à participação de contribuidores com diferentes entendimentos de valores humanos no processo de Design de um artefato. Assim, os valores humanos seriam definidos a partir de restrições e melhorias providenciadas pelo artefato, dado que essa tecnologia será construída com as necessidades de usuários (humanos) em mente e também será influenciada pelos valores humanos dos contribuidores, destacando certos valores mais do que outros.

Para Pereira, Baranauskas e Liu (2018), valores são "algo que tem importância para alguém dentro de um certo contexto ou com alguma capacidade", de forma que no contexto de valores humanos, esse "alguém" é o ser humano. Os autores também discutem o conceito de regras ideais, que são "tipos especiais de normas que comunicam valores, especificam características relacionadas a valores a ajudam a traduzí-las para uma estrutura mais formal". Para os autores, essas regras são úteis para IHC devido à possibilidade de utilizá-las como uma forma de unir um contexto social complexo, envolvendo a cultura e valores de diversos contribuidores, por exemplo, e os artefatos projetados, que promovem essas culturas e valores (SALGADO, DE SOUZA, LEITÃO., 2011; SALGADO, LEITÃO, DE SOUZA, 2012).

Esses pontos são essenciais para esse trabalho, que tenta analisar violações de valores humanos derivadas tanto do campo de Design quanto do campo de IA, por meio do estudo do Twitter. Assim, as questões sobre direitos humanos serão discutidas a partir de uma visão de design voltado à promoção de valores humanos.

3. TRABALHOS RELACIONADOS

Ao longo dos anos, foram feitos pesquisas e estudos sobre como as redes sociais empregam práticas éticas em seus serviços e produtos. A seguir, será feita a revisão de alguns trabalhos sobre o assunto.

Um dos estudos feitos foi o de Woolley e Howard (2019). Nele, são analisadas as contribuições das mídias sociais para formar opiniões políticas de indivíduos. Comentam que grandes redes sociais como o Twitter têm dificuldade em gerenciar o controle político dentro de seu produto principal, devido ao que chamam de câmaras de eco (ambientes que contém opiniões polarizadas) criadas a partir dos algoritmos da rede. Graças a esses algoritmos, seria possível que grupos pequenos ampliassem a disseminação de informações, verdadeiras ou falsas, em larga escala. Além disso, notam também que a presença de robôs (conhecidos como *bots)* na rede contribui e facilita ainda mais essa disseminação, sendo comumente usados em campanhas políticas para construir uma determinada opinião sobre um candidato à eleição. Embora os autores analisem os efeitos da propaganda computacional, o escopo das redes analisadas é mais amplo e o tópico escolhido é único, carecendo de uma compreensão das estratégias de comunicação sobre valores éticos por parte do Twitter e um definições de ética mais delineadas e bem definidas, embora o livro faça, indiretamente, uma análise sobre privacidade e segurança.

Há também o trabalho de Burr e Cristianini (2019) que revisa diversos estudos sobre como os algoritmos de aprendizado de máquina podem inferir características e estados mentais e emocionais a partir de um conjunto de dados relacionados a indivíduos. Nele, os autores levantam questões sobre casos como esses, e para além disso, casos em que algoritmos conseguem agrupar indivíduos baseados em características em comum. Eles notam que esses tipos de tecnologia estão se desenvolvendo descoordenadamente, prejudicando usuários. Outro argumento que levantam é de que diferentes usos dessas tecnologias levam a diferentes vieses e, portanto, diferentes preocupações éticas e quando o usuário não tem conhecimento desse risco, ele torna um perigo alarmante e grande o suficiente para superar regulações e leis sobre privacidade. Em suma, trabalho se preocupa mais em entender os possíveis efeitos colaterais éticos do uso de IA para automatizar testes psicológicos (como por exemplo, negar um diagnóstico de depressão pois um usuário de redes sociais interage com ou publica conteúdos positivos) e não em avaliar eticamente um sistema que utiliza algoritmos desse tipo, embora confirme e ressalta as preocupações geradas por eles.

Auernhammer (2020) também discute sobre os efeitos da IA. Nesse trabalho, o autor ressalta que práticas de design são importantes para o desenvolvimento de sistemas que utilizam IA. O autor mostra que muitas políticas ou guias para o desenvolvimento ético de IA são insuficientes e pouco efetivas para problemas complexos do mundo real, devido a fatores como desenvolvimento mais devagar de leis que tecnologia, diferenças culturais de definição de ética e falta de previsibilidade dos efeitos da IA.

Auernhammer exibe, analisa e compara alguns tipos de práticas de design, incluindo o Design Thinking, (BROWN, 2018) comentando que ainda há pouca relação ou colaboração entre os campos de Design e IA. Assim, o trabalho dele se difere deste devido a não comentar sobre redes sociais especificamente e nem sobre o caráter psicológico como o trabalho anterior, além de abordar ética de uma forma geral. Entretanto, percebe-se que também motiva uma conexão entre o âmbito de Design e IA para solucionar problemas éticos.

Em seu estudo, Harris (2016) discute como a tecnologia é capaz de aproveitar das fraquezas da mente humana. O autor dá diversos exemplos de tecnologias que são capazes de induzir os seus usuários a certos hábitos ou atitudes, como percorrer a linha de tempo por tempo indeterminado (em uma analogia com caça-níqueis), obrigar um usuário a interagir com outras funcionalidades do sistema ao invés de oferecer opções de fácil uso para ele cumprir a tarefa desejada (como o Twitter obriga um usuário à ver sua linha de tempo mesmo que ele só queira publicar algo) e até mesmo incentivando que outras pessoas comentem uma nova foto de perfil nas redes sociais. O Facebook, por exemplo, exibe a nova foto com mais prioridade na linha de tempo de amigos para que possam interagir mais na rede. Resumindo, o artigo evidencia e explica algumas práticas antiéticas que a tecnologia tem, mas não as relaciona especificamente com princípios éticos ou com o Twitter. Além disso, também não discute algoritmos de IA ou *Big Data*, que, como já dito anteriormente, são responsáveis por grandes problemas éticos em redes sociais, se restringindo apenas ao design de experiências de aplicações, incluindo redes sociais.

Outro trabalho que faz análises éticas é o de Jones (2017). Nele, o autor analisa a construção do design de serviços de redes sociais, criticando que comumente infringem princípios éticos. Para tanto, Jones parte de uma contextualização socioeconômica das redes sociais e seus modelos de negócio, trazendo à tona a relação entre redes sociais, publicidade, governo e corporações, todas de alguma forma reduzindo a privacidade do usuário e o controle que ele tem sobre si. Embora Jones faça um excelente trabalho em expor a criticidade das atitudes dos projetistas de serviços de redes sociais, analisando os padrões de design que limitam a privacidade; e contextualizando-as em uma esfera socioeconômica onde as TICs

possam ser usadas para controlar indivíduos, o autor não faz menção ao Twitter (restringindo-se majoritariamente ao Facebook) e foca principalmente em aspectos éticos de transparência e privacidade, relacionando ambas.

Um trabalho similar ao anterior é o de Light e McGrath (2010). Eles realizam uma análise descritiva do Facebook a partir da observação de usuários, extraindo possíveis preocupações com a privacidade e segurança decorrentes do uso da rede social. A ética e a moral são analisados durante: a criação de perfil, o preenchimento de informações adicionais após a criação do perfil, o compartilhamento de conteúdo e publicações de atividade, encontrando complicações éticas em todas elas, a exemplo de cadastrar aplicações no perfil do Facebook, onde os dados do usuário são compartilhados com entidades desconhecidas e com limite também desconhecido.

Light e McGrath deixam claro que tem interesse em entender as crenças e valores sobre moral e ética de determinados agentes, ao invés de estabelecerem princípios de antemão para só então fazerem a análise. Essa é uma das maiores diferenças entre o trabalho deles e este, já que aqui escolhemos princípios antes de realizar uma análise. O quadro 1 indica um resumo com os trabalhos analisados neste capítulo:

Quadro 1 – Lições aprendidas com os trabalhos relacionados

Estudo	**Foco do estudo**	**Principais resultados**	**Relevância**
Woolley e Howard (2019)	Influência de opiniões políticas por meio de redes sociais.	É possível utilizar *bots*, câmaras de ecos nas redes sociais para realizar campanhas políticas dentro de redes sociais.	O estudo mostra que ocorre propaganda computacional em redes sociais, mas não considera princípios éticos e nem as estratégias de comunicação do Twitter.
Burr e Cristianini (2019)	Inferência de estado emocional e mental de indivíduos.	Existem métodos que utilizam IA para inferir o estado emocional e mental de indivíduos a partir de um conjunto de dados, colocando-os em risco.	Os autores demonstram preocupações em automatizar testes psicológicos, utilizando IA para

			inferir emoções dos participantes, evidenciando um problema de segurança, embora não discuta esse processo de inferência para redes sociais.
Auernhammer (2020)	Relação entre IA e Design.	É importante e possível utilizar métodos de Design, como Design Thinking, no desenvolvimento de IAs.	Neste estudo, há uma motivação para utilizar práticas de Design para desenvolver modelos de IA mais éticos, mostrando uma relação entre esses dois campos.
Harris (2016)	Influência da tecnologia no comportamento humano.	Há padrões em redes sociais, por exemplo, que influenciam sutilmente o usuário a ter certos comportamentos, como navegar em uma rede social por tempo indeterminado e ininterruptamente.	O autor mostra que as redes sociais possuem diversas práticas antiéticas em seu design, motivando uma análise mais detalhada, mas não comenta sobre a influência de IA ou *Big Data* nesse processo.
Jones (2017)	Relação entre aspectos socioeconômicos individuais e redes sociais	Governos e corporações detêm dados de usuários de redes sociais ou que podem comprometer a privacidade	O autor critica as organizações que gerenciam redes sociais, dizendo que elas implementam

		do usuário e seu controle sobre si mesmo.	práticas que limitam a privacidade dos usuários, com foco nos princípios éticos de privacidade e transparência.
Light e McGrath (2010)	Entendimento de crenças e valores éticos dos projetistas de uma aplicação	A partir da análise descritiva de uma aplicação e observação dos usuários em um contexto de uso dela, é possível determinar as crenças e valores éticos dos projetistas de uma aplicação.	Os autores mostram que é possível entender as crenças e valores éticos dos projetistas de uma aplicação através da sua análise, permitindo que sejam identificados problemas éticos nela.

4. ESTUDOS REALIZADOS

Este capítulo apresenta os passos para revelar potenciais problemas na comunicabilidade da ética pelo Twitter; entender se as pessoas usuárias do Twitter enfrentam ou percebem problemas éticos durante o uso cotidiano; e entender a percepção dessas pessoas sobre Ética.

Conforme dito na introdução (seção 1.2), este trabalho teve como objeto de estudos o Twitter, que é introduzindo na seção 4.1 para a contextualização do estudo. A seção 4.2 descreve a inspeção preliminar realizada para escolher princípios éticos para IA. Na seção 4.3 é aplicado o MIS, orientado pelos princípios éticos para IA escolhidos na seção 4.2. Na seção 4.4, são apresentadas e discutidas as respostas da pesquisa com usuários do Twitter. Finalmente, na seção 4.5, discute-se os resultados dos estudos feitos nas seções 4.3 e 4.4.

4.1. A rede social Twitter

A ferramenta estudada neste trabalho é o Twitter, uma rede social onde as interações entre usuários ocorrem através de publicações pequenas (*tweets*), como mostra a Figura 1, de no máximo 280 caracteres, que podem conter imagens, vídeos ou enquetes.

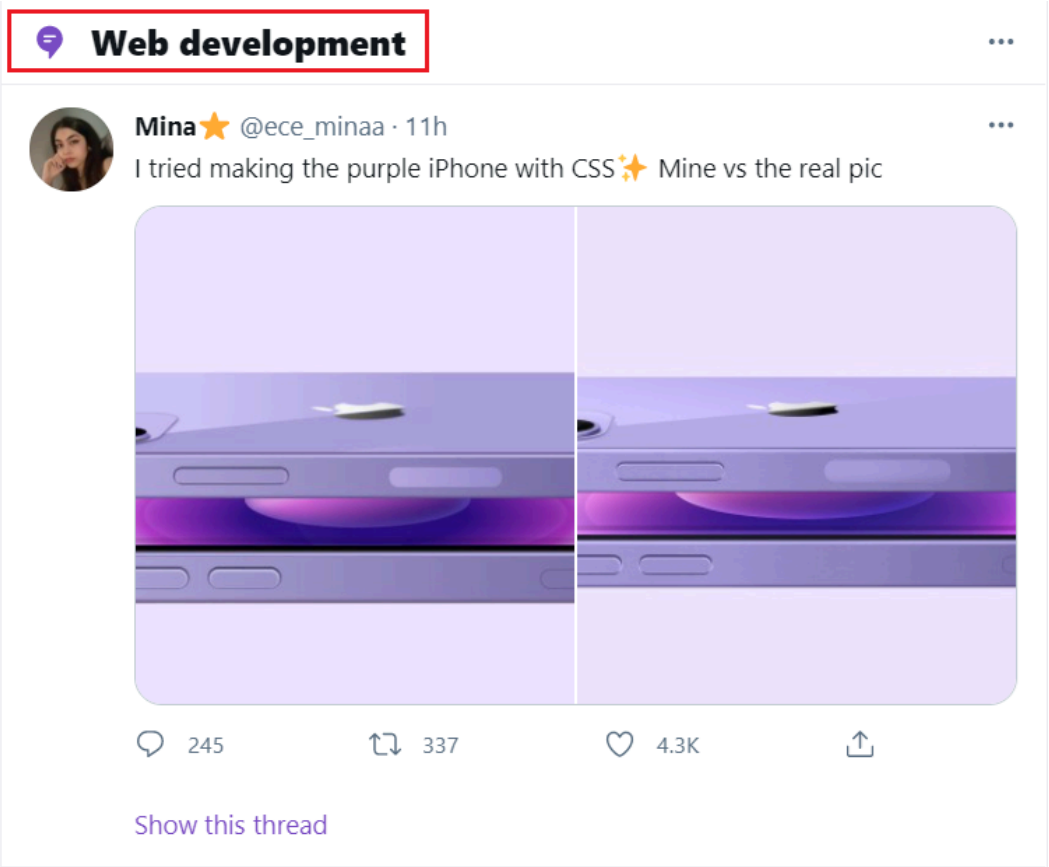
Usuários podem seguir outros perfis pelos quais se interessam, como celebridades e pessoas públicas, para acompanhar suas publicações por meio de uma linha de tempo. É possível interagir com as publicações através de curtidas ♡, compartilhamento (retuíte) e comentários, como mostra a Figura 1.

Figura 1 - Um exemplo de publicação no Twitter. Signos, da esquerda para direita: comentar, compartilhar e curtida

Fonte: https://twitter.com/JustinTrudeau/status/1386362640897912837

Analogamente, também é possível seguir tópicos de interesse (vide Figura 2), como esportes, política e ciência, e acompanhar notícias recentes através da funcionalidade "Assuntos do Momento", que mostra as publicações e tópicos mais populares naquele momento.

Figura 2 - Exemplo de tópico de interesse (*Web Development*), destacado com borda vermelha.



Fonte: https://twitter.com/explore

4.2.Inspeção preliminar do Twitter e escolha dos princípios éticos para IA

Visto os potenciais problemas éticos que podem acontecer em redes sociais, investigamos possíveis cenários que violam os princípios éticos para IA dentro do Twitter. Para tanto, foi realizada uma inspeção da tecnologia utilizando o trabalho de Fjeld et al. (2020). A pesquisa descreve uma gama de princípios éticos que são impactados pelo uso da IA e possibilitou uma análise da aplicação através das lentes desses princípios.

Neste processo preliminar, encontrou-se reportagens, artigos e seções do site que valorizam tais princípios. Além disso, foram identificadas outras evidências que impactam negativamente os princípios éticos de Privacidade, Segurança, Promoção de Valores Humanos, Não-Discriminação, Transparência e Explicabilidade descritos por Fjeld et al. (2020).

Quanto à **Privacidade e Segurança**, uma das reportagens, feita por Reed (2020), denunciava que alguns perfis de pessoas públicas importantes (como o ex-presidente dos Estados Unidos da América Barack Obama) foram invadidos por *hackers*. Com essas ações,

eles foram capazes de publicar postagens para roubar quantias de *Bitcoin* dos seguidores dos perfis através de discursos falsos que prometiam devolver a quantia doada. Além disso, os invasores foram capazes de baixar todos os dados dos perfis a partir de uma ferramenta disponibilizada pelo próprio Twitter. Isso evidencia uma falha de privacidade e também de segurança, já que houve roubo de dados privados dos perfis e roubo de dinheiro dos seguidores.

Outra evidência é encontrada em um artigo[8] na Central de Ajuda do Twitter, e revela fragilidades da **Transparência e Explicabilidade** dos algoritmos de recomendação do Twitter nas explicações da Central de Ajuda do site. Embora as explicações descrevam como funcionam um determinado trecho do Twitter relativo aos "assuntos do momento" mais populares na rede, essa descrição é breve. Não há referências internas ou externas para uma explicação profunda do funcionamento dos algoritmos que são utilizados para construir essa funcionalidade e, portanto, que informe sobre os impactos indiretos que um perfil terá ao interagir com eles. A pouca transparência dificulta, por exemplo, a análise dos impactos aos princípios éticos decorrentes do algoritmo e até a criação de um modelo de governança ou políticas para manter uma boa ética.

O Twitter contém funcionalidades que favorecem o princípio ético da N**ão-Discriminação**, igualdade e respeito, como permitir que sejam adicionadas descrições nas imagens publicadas no site para que pessoas com deficiências visuais consigam interpretá-las. Na Central de Ajuda (2020), são exibidas políticas firmes sobre esse quesito, como sistemas de punição e suspensão de perfis que participam desses atos, visando construir uma comunidade não-discriminatória. Entretanto, a pesquisa de ElSherief et al. (2018) oferece uma amostragem e análise de vários discursos de ódio encontrados na plataforma, apontando que as políticas construídas pelo Twitter não estão sendo tão efetivas quanto deveriam. Portanto, o princípio ético de não-discriminação ainda não foi suficientemente garantido no site.

Quanto à **Promoção de Valores Humanos**, foi encontrado um contraste quanto à propagação de ódio no Twitter. Vincent (2016) denuncia um experimento da Microsoft que ocorreu na rede social em 2016: pesquisadores criaram um robô de conversação chamado "Tay", que aprendia através da interação com usuários do Twitter. Apesar de terem tido boas intenções, alguns perfis falavam discursos de ódio, como discursos xenofóbicos, racistas,

[8] Disponível em: https://help.twitter.com/pt/using-twitter/twitter-trending-faqs. Acesso em 6 de dez. de 2020

misóginos, ultrapassando o filtro implementado pelos pesquisadores para prevenir isso. Ao invés de promover valores humanos, os usuários da plataforma a utilizaram para promover conflitos e disseminar discursos de ódio. Após algum tempo, o robô foi removido da plataforma, assim como algumas publicações relacionadas a essas ocorrências. Mesmo assim, esse exemplo revela uma fragilidade na possibilidade ética do Twitter em promover valores humanos.

Dadas as evidências encontradas na inspeção preliminar, delineamos e escolhemos alguns princípios éticos que carecem de melhorias no site da rede social para serem avaliados com mais profundidade: **Privacidade**, **Segurança**, **Promoção de Valores Humanos**, **Não-Discriminação**, **Transparência** e **Explicabilidade**. Esta escolha teve como base os problemas éticos encontrados na inspeção preliminar, classificando-os de acordo com as definições no trabalho de Fjeld et al. (2020).

Neste trabalho, os princípios de **Privacidade** e **Segurança** e serão analisados juntos, assim como os de **Transparência** e **Explicabilidade**, devido ao entendimento de que estão relacionados, como já conceituado nas seções 2.2 e 2.3.

4.3.Inspeção Semiótica do Twitter

Para avaliar com mais profundidade o Twitter com relação aos cuidados éticos que a ferramenta tem *by design*, foi utilizado o Método de Inspeção Semiótica (DE SOUZA et al., 2006; DE SOUZA & LEITÃO, 2009). Esse método tem como objetivo avaliar a comunicabilidade de uma interface, que é a capacidade da interface em comunicar aos usuários quem os designers da ferramenta acham que é(são) a(s) pessoa(s) usuária(s) principais, como devem utilizá-la, de que formas preferenciais e porquê, de forma a atender seus respectivos objetivos. O método permite, então, que um especialista em IHC inspecione a interface em questão para compreender a lógica por trás de seu design. Isso ocorre através da (re)construção de metamensagem (o que o designer queria comunicar ao usuário) com a análise dos signos metalinguísticos, estáticos e dinâmicos (DE SOUZA et al, 2006).

Nesse contexto, Souza & Leitão adotam a definição de signos de Peirce e Houser (1998), o qual os define como qualquer coisa que tenha significado para alguém. As autoras classificam os signos como estáticos se não possuírem mudança de estado; dinâmicos caso possuam mudanças de estado ao longo do tempo; e metalinguísticos caso sejam utilizados para explicar ou esclarecer outro tipo de signo (DE SOUZA & LEITÃO, 2009).

Como descrito por de Souza e Leitão (2009), o método é dividido em 5 etapas sequenciais para atingir o seu objetivo. Inicialmente, deve-se definir o perfil do usuário que realizará a tarefa desejada, assim como escolher as partes da interface envolvidas nesta, e escrever o cenário de inspeção. Em seguida, inspeciona-se a interface, analisando os signos encontrados, seguindo a ordem: metalinguísticos, estáticos e dinâmicos.

Continuando, para cada signo, reconstrói-se a metamensagem dele, buscando entender e preencher o modelo de metacomunicação que diz:

"*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"*.

No final, realiza-se um contraste entre as três metamensagens reconstruídas, buscando por redundâncias, (in)consistências e como a carga de comunicação é distribuída entre os signos, avaliando também a qualidade das metamensagem.

Neste trabalho, para avaliar cada princípio ético, foi elaborado um cenário de inspeção específico para cada, assim como foram selecionadas partes da interface, utilizando o mesmo perfil de usuário para todo o respectivo cenário de inspeção. Assim, foram realizadas, ao todo, quatro execuções individuais e não relacionadas do método.

O perfil de usuário elaborado foi uma pessoa jovem, na faixa de 18-25 anos de idade, com familiaridade razoável com computadores e smartphones, que utiliza o Twitter frequentemente durante o dia. A seguir, exibe-se os resultados de cada etapa da aplicação do MIS para os princípios selecionados:

4.3.1. Privacidade e Segurança

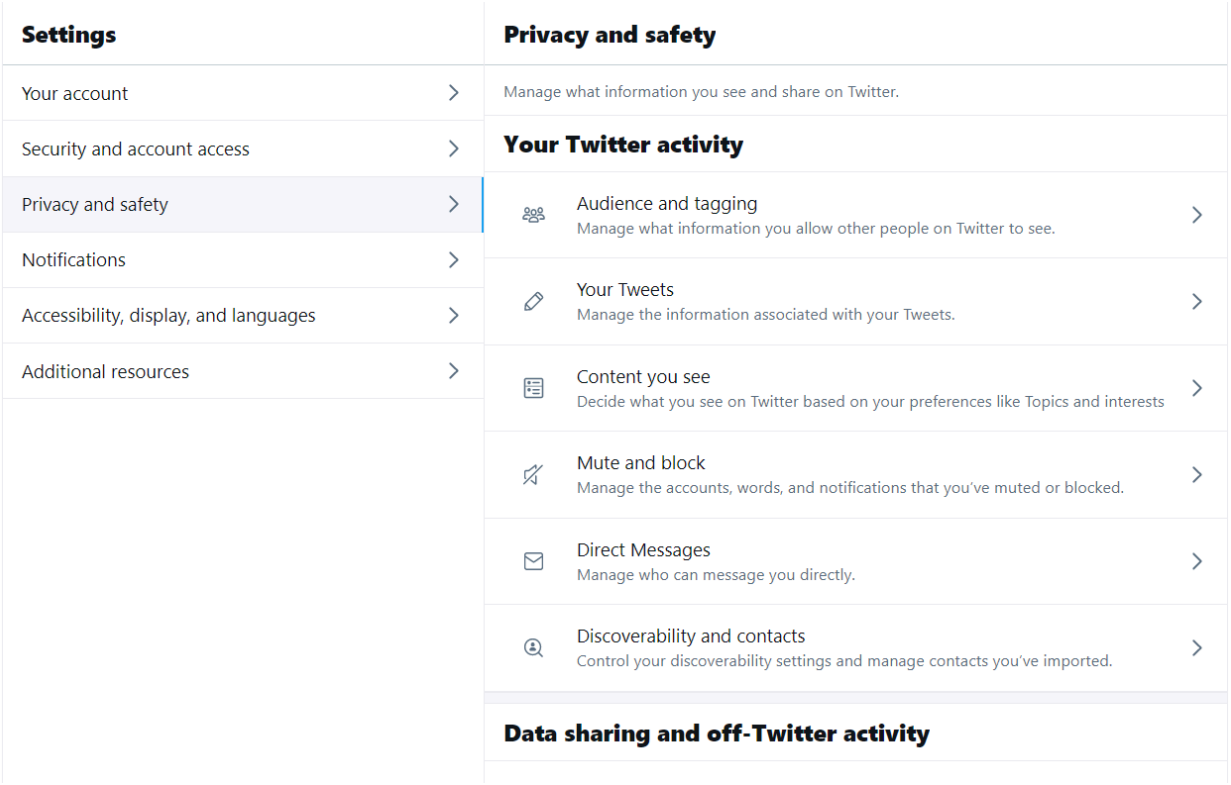
O cenário elaborado para analisar esses princípios foi: "*Beatriz, uma jovem universitária de Direito, está estudando para um processo seletivo de estágio na área de Direito Digital. Por ser iniciante no assunto, começou a estudar um dos tópicos que as empresas costumam demonstrar preocupação: proteção dos dados de usuários. Após ler alguns estudos sobre a LGPD, decidiu desabilitar as opções de privacidade em seu* Twitter *relacionadas a compartilhamento de seus dados, assim como fechar a visibilidade de seu perfil e apagar os seus dados.*"

Serão realizadas, portanto, três tarefas: desabilitar o compartilhamento de dados, fechar a visibilidade de seu perfil e apagar os dados do perfil armazenados na plataforma. No

início, o menu de configurações gerais é acessado na subseção "Privacidade e Segurança", como mostra a Figura 3.

Então, escolhe-se as opções correspondentes a cada tarefa para cumprí-las. Para a tarefa de fechar a visibilidade de seu perfil, a pessoa usuária deve selecionar a opção "*Protect your Tweets*" dentro do menu "*Audience and Tagging"*. Para desabilitar o compartilhamento de dados, a pessoa usuária deve entrar no menu de "*Data sharing with business partners*", na seção "*Data sharing and off-Twitter activity*" e desabilitar a opção "*Allow additional information sharing with business partners*"

Figura 3 - Menu de configurações gerais do Twitter

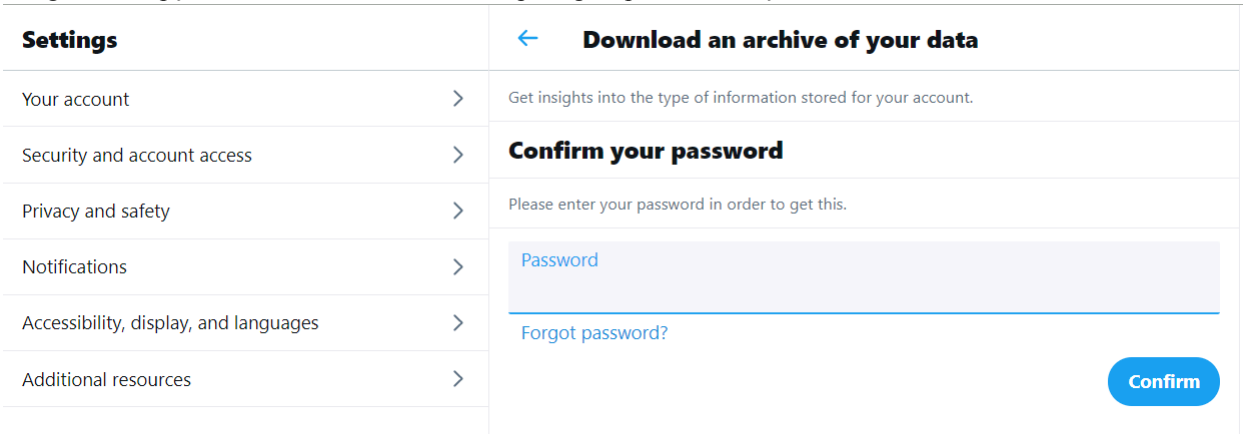


Fonte: Configurações do Perfil no Twitter[9]

É possível perceber que a opção para gerenciar a privacidade e segurança do perfil, que contém as opções para resolver as duas primeiras tarefas, é bem explícita, mostrando que os designers da interface se importam com esses princípios. Entretanto, essa opção não está disponível diretamente a partir da página principal, então a preocupação não é tão grande assim. Além disso, cada subseção possui explicações para apoiar o usuário ou até mesmo links para a Central de Ajuda do site, provavelmente para ajudar usuários que não têm muita familiaridade com a plataforma ou esses tipos de tarefas.

[9] Disponível em: https://twitter.com/settings/privacy_and_safety. Acesso em 6 de dez. de 2020.

Figura 4 – Opção de baixar os dados da conta protegida por confirmação da senha da conta



Fonte: Página de configuração no Twitter[10]

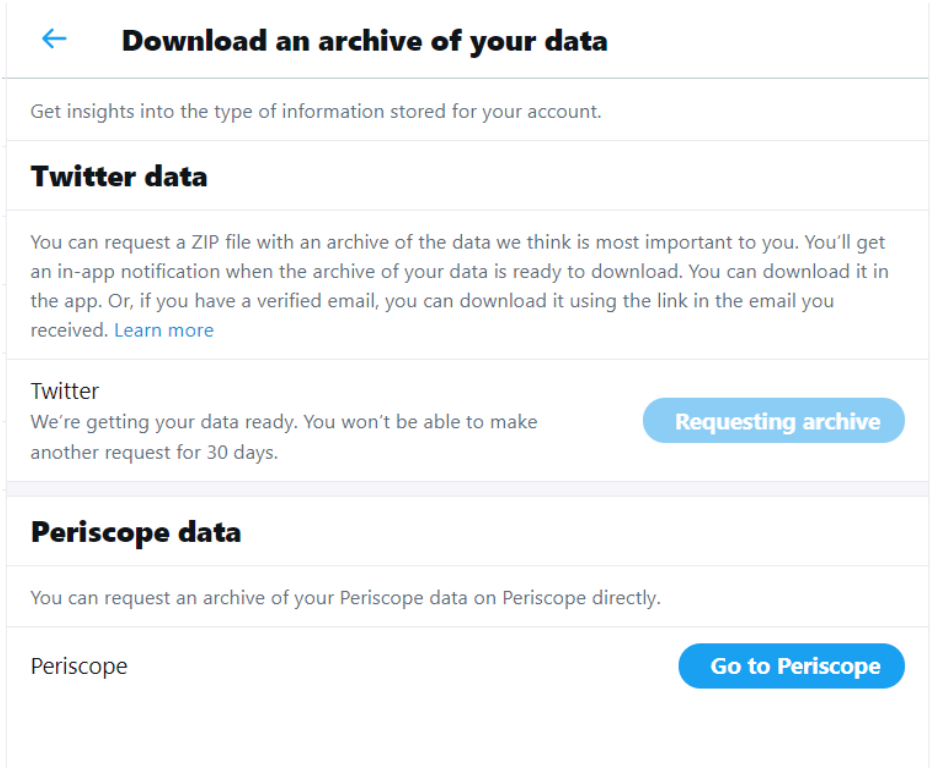
Uma prática de privacidade e segurança por parte dos designers foi permitir que o usuário baixasse seus dados e eles só podem ser baixados após inserir a senha corretamente, a exemplo da interface na Figura 4. Assim, só quem sabe a sua senha pode acessar seus dados. Ainda, caso você tenha esquecido sua senha, é possível redefini-la a partir daquela mesma tela, sem que seja necessário sair da sessão atual.

Entretanto, vale notar alguns pontos negativos quanto a essas abordagens. Primeiramente, qualquer pessoa que possua a sua senha terá acesso aos seus dados. Embora o Twitter forneça métodos mais seguros de autenticação utilizando autenticação de dois-fatores, essa opção não se apresentou disponível para a baixar os dados do perfil. Isso significa que algum invasor que possua a sua senha conseguirá acessar todos os seus dados, configurando uma espécie de autenticação falsa.

Segundo, não há nenhuma notificação que diz que o processo só pode ser feito a cada 30 dias, algo que só é descoberto após efetivamente baixar os dados. Essa foi uma grave falha do Twitter, pois não há mecanismos para confirmar se você realmente quer baixar os seus dados antes da operação iniciar, o que permite enganos e erros por parte do usuário, como ocorre na Figura 5. Além disso, não é possível interromper o processo, agravando ainda mais a falha.

[10] Disponível em: https://twitter.com/settings/download_your_data. Acesso em 6 de dez. de 2020.

Figura 5 – Botão de baixar os dados da conta desativado por 30 dias após clique



Fonte: Baixar seus dados no Twitter[11]

Por fim, é possível reconstruir a metamensagem de cada signo. Iniciando com os metalinguísticos, têm-se:

- **Quem é o usuário:**

  "*Olá Beatriz. Entendemos que você usa o* Twitter *e possui certas preocupações sobre sua privacidade e segurança. Entendemos que você é uma usuária que pode ter dificuldade em entender textos e títulos curtos.*"
- **O que ele deseja fazer:**

  "Você quer entender sobre as configurações de privacidade disponíveis."
- **De que formas prefere fazer:**

  "Através de informações e explicações expostas diretamente na interface."
- **Qual sistema foi construído:**

  "Para isso, nós, designers do Twitter*, projetamos este sistema. Nele, as informações e opções dos menus possuem breves explicações.*"

[11] Disponível em: https://twitter.com/settings/download_your_data. Acesso em 6 de dez. de 2020.

- **De que formas deve ser utilizado:**

  "As explicações podem ser vistas diretamente na interface de forma breve, aprofundando à medida que você navega pelas seções."

- **Para atingir quais objetivos:**

  "*Entender a seção onde você se encontra, as opções e ações disponíveis assim como suas consequências.*

Seguindo para os signos estáticos, temos a seguinte mensagem:

- **Quem é o usuário:**

  "O*lá Beatriz. Entendemos que você usa o* Twitter *e possui certas preocupações sobre sua privacidade e segurança.*"

- **O que ele deseja fazer:**

  "Você quer gerenciar as configurações de privacidade e segurança dos seus dados e da sua conta, além de garantir a segurança do acesso aos seus dados."

- **De que formas prefere fazer:**

  "Você é uma usuária que prefere realizar suas tarefas fácil e rapidamente e que gosta de explicações breves com possibilidade de entender mais sobre os assuntos relacionados."

- **Qual sistema foi construído:**

  "Para isso, nós, designers do Twitter, *projetamos este sistema, com informações devidamente categorizadas e divididas em subseções, caixas de seleção para te dar total controle sobre o gerenciamento e botões para facilitar a sua interação com o site e uma opção para baixar os seus dados no Twitter. Sobre segurança, acreditamos que embora você queria garantir seu acesso individual aos dados, não há formas de garantir acesso aos seus dados através de autenticação de dois fatores e após solicitado o download deles, não é possível reverter o processo."*

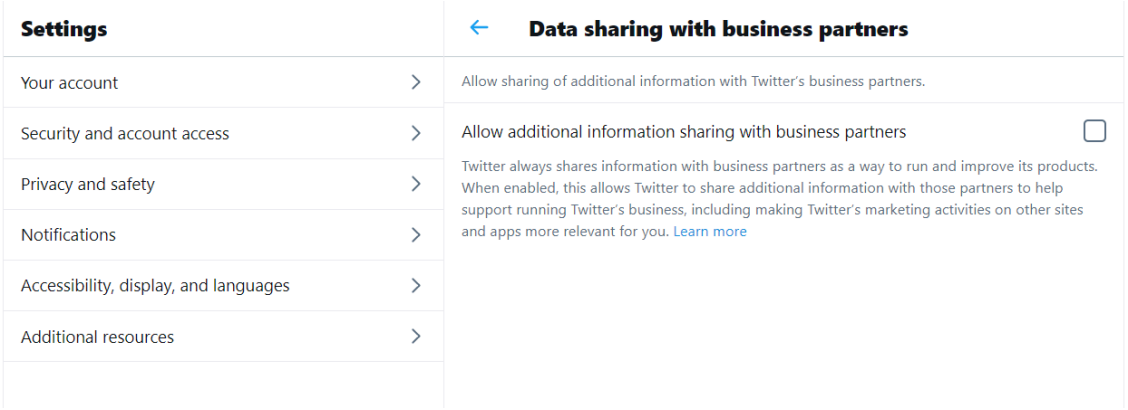
- **De que formas deve ser utilizado:**

  "As explicações podem ser vistas próximas as subseções, opções ou ações. Para gerenciar a sua privacidade, você pode marcar ou desmarcar caixas de seleção que configuram o compartilhamento de dados ou a visibilidade de seu perfil. Para baixar os seus dados, você pode clicar no botão para solicitá-los."

- **Para atingir quais objetivos:**

"Com esse sistema, você conseguirá interromper o compartilhamento de dados com os parceiros do Twitter e limitar a visibilidade de seu perfil, entender as opções e seções disponíveis nas configurações e baixar os dados de sua conta que achamos importantes."

Figura 6 – Caixa de seleção para habilitar ou desabilitar o compartilhamento de dados da conta



Fonte: Compartilhamento de dados com parceiros, no Twitter[12]

Reconstruindo a metamensagem dos signos dinâmicos, temos a seguinte metamensagem:

- **Quem é o usuário:**
  "Olá Beatriz. Entendemos que você usa o Twitter *e possui certas preocupações sobre sua privacidade e segurança. Você é uma usuária que não é muito insegura em tomar decisões."*
- **O que ele deseja fazer:**
  "Você deseja um sistema que te informe quando está tomando uma ação e em que local do processo você se encontra, de forma clara e rápida."
- **De que formas prefere fazer:**
  "Você prefere de indicadores sutis sobre o que pode ou não fazer. Além disso, você também gosta de acessar opções de maneira fácil e rápida."

- **Qual sistema foi construído:**

[12] Disponível em: https://twitter.com/settings/data_sharing_with_business_partners. Acesso em 6 de dez. de 2020.

"Visto isso, projetamos esse sistema onde as partes interativas mudam de cor ou são realçadas ao passar o mouse por cima delas ou clicar quando necessário."

- **De que formas deve ser utilizado:**
  "Você deve passar o mouse por cima das opções e botões para que identifique se eles estão habilitados ou não para alterações ou ações."
- **Para atingir quais objetivos:**
  "Entender se uma opção relacionada a privacidade e segurança pode ser alterada ou não."

Em suma, percebe-se que o Twitter possui estratégias de comunicação que enfatizam a compreensão das ações a serem tomadas e garantir o máximo de controle para o usuário. Os projetistas potencialmente atingirão isso através de caixas de seleção, como na Figura 6, que possuem estados alternantes entre habilitar e desabilitar uma opção; através de ícones para facilitar a identificação e navegação entre as páginas, como na Figura 3; e através de explicações localizadas próximas às opções que um usuário pode selecionar e links para seções de ajuda em um site externo, algo que é mostrado nas Figuras 3 e 4. Essas estratégias facilitam o uso do site da aplicação, embora existam algumas dissonâncias no design a serem melhoradas.

Primeiramente, há uma enorme redundância entre os signos metalinguísticos e estáticos, pois os títulos das seções são quase sempre seguidos de uma explicação. Para usuários inexperientes, isso se torna uma explicação auxiliar; mas para usuários mais experientes, isso pode ser informação em excesso.

Adicionalmente, a ação de baixar os dados do perfil gera problema de contraste. Após clicar no botão para baixar o arquivo, o botão é desabilitado até que os arquivos estejam prontos para serem baixados. Isso não foi avisado com antecedência, e não se diz quando os arquivos estarão prontos, o que pode levar uma quantidade arbitrária de tempo. Ademais, essa ação é irreversível, o que é diferente do design geral de outras tarefas.

Por último, percebe-se então que a terceira tarefa não pôde ser completada. Embora o usuário da rede tenha direitos sobre seus dados, não é possível removê-los, pois o Twitter os armazena indeterminadamente. Existem dois pontos que valem uma discussão: o fato do Twitter não permitir a remoção dos dados (pelo menos não de forma direta); e o armazenamento por tempo indeterminado.

Entende-se que para uma rede social, não remover dados é importante para melhorar a experiência dos seus usuários, seja de forma direta através de melhor capacidade de

perfilamento pelos algoritmos, como explicado por Kasai, Yusof e Clarke (2016); ou de forma indireta através da melhoria em tomada de decisões suportada pela inteligência de negócio, como diz Pratt (2019). Entretanto, o gerenciamento de privacidade e segurança de um usuário é dificultado.

Sobre a ação de baixar os dados do perfil, os projetistas do Twitter comunicam que os dados disponíveis para baixar são apenas os que eles consideram importantes, sem mencionar se esses dados são recentes ou antigos, o que gera a dúvida: **Por quanto tempo seus dados são armazenados**? A ausência de resposta abre espaço para que sejam feitas análises em seu perfil por parceiros da plataforma utilizando dados antigos, que não mais relevantes, o que prejudica a experiência do usuário já que contraria o benefício de mantê-los. Esses dois pontos denunciam ainda mais a necessidade de melhorias relacionadas à privacidade e segurança de usuários do Twitter.

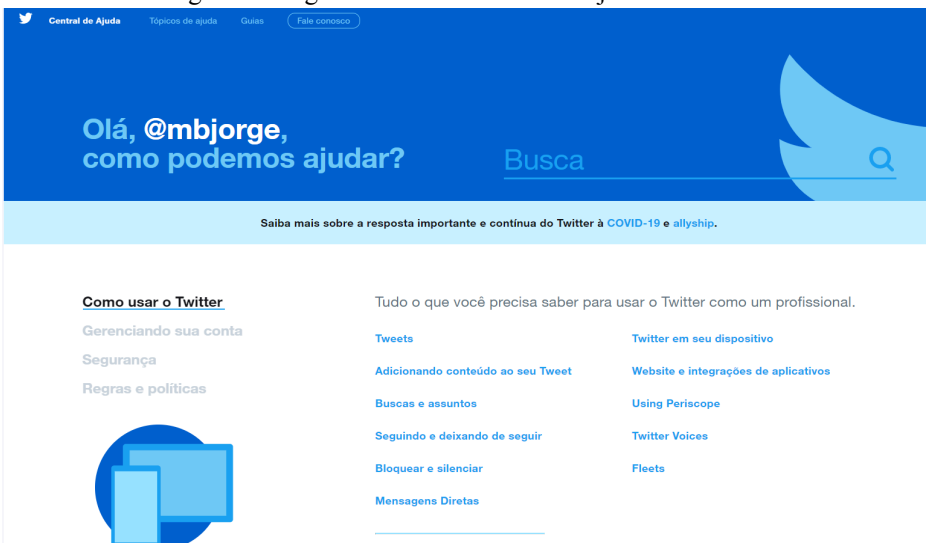
Tudo isso contribuiu para mostrar que o Twitter, embora preocupado com privacidade e segurança, ainda precise melhorar em certos aspectos, visto que a tarefa de remover os dados não é completável, inibindo o controle do usuário sobre seus dados; não há explicações detalhadas nem alertas (como outras categorias possuem) sobre baixar os dados do perfil; e os dados são armazenados por tempo indeterminado. Em outras palavras, enquanto em certos tópicos o Twitter explica demais sobre o assunto, quando se trata de dados e privacidade, ele se mostrou pouco ou nada explicativo.

4.3.2. Transparência e Explicabilidade

Para este princípio, o cenário elaborado foi: "*Jorge é um jovem que adora navegar pelas redes sociais. Todos os dias ele entra no* Twitter *a cada 30 minutos, aproximadamente, para ver os assuntos do momento e acompanhar os perfis que segue através de sua linha de tempo. Após ver alguns tópicos que achou estranho, ele decidiu buscar mais informações sobre por que isso aconteceu, a partir da Central de Ajuda do site. Em seguida, ao se deparar com tweets que não são de seu interesse, ele marca-o devidamente como 'Não interessado'*".

Neste cenário, serão inspecionadas duas tarefas: entender os fatores que levam um conteúdo a ser mostrado para o usuário; e demonstrar desinteresse sobre uma publicação. Inicialmente, abre-se a Central de Ajuda do Twitter, através de uma opção no submenu inicial do site. A Central conta com artigos explicativos sobre certas funcionalidades da plataforma (nota-se o caráter metalinguístico desse tipo de signo). Nela, é possível buscar por termos para facilitar o encontro do tópico de interesse (Figura 7).

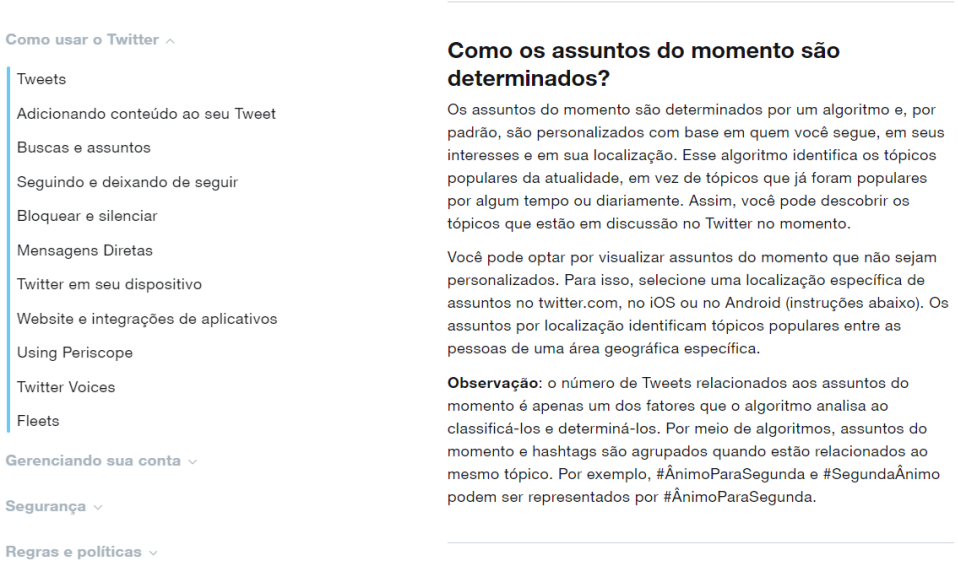
Figura 7 - Página inicial da Central de Ajuda do Twitter



Fonte: Central de Ajuda do Twitter[13]

Para cumprir a primeira tarefa, busca-se por "assuntos do momento" e "tópicos", selecionando os artigos relacionados dentre os resultados (Figura 8 e 9, respectivamente).

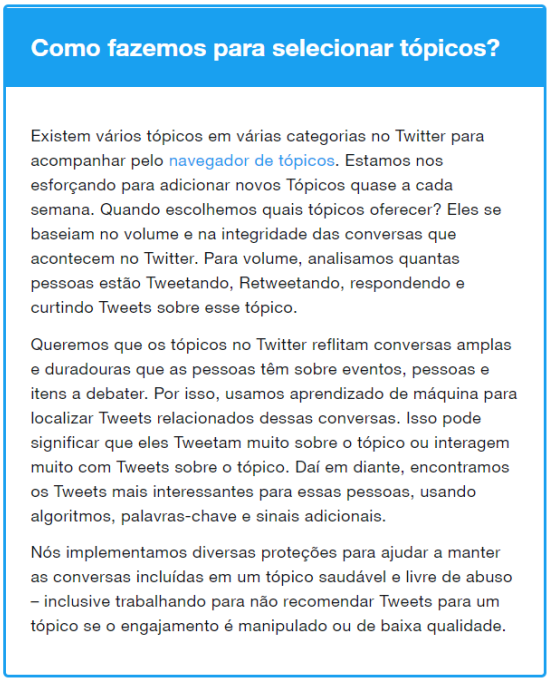
Figura 8 - Trecho de artigo da seção de ajuda explicando como funcionam os assuntos do momento



[13] Disponível em: https://help.twitter.com/pt. Acesso em 6 de dez. de 2020.

Fonte: Central de ajuda do Twitter[14]

Figura 9 - Trecho de artigo da seção de ajuda explicando como os tópicos são selecionados



Como fazemos para selecionar tópicos?

Existem vários tópicos em várias categorias no Twitter para acompanhar pelo navegador de tópicos. Estamos nos esforçando para adicionar novos Tópicos quase a cada semana. Quando escolhemos quais tópicos oferecer? Eles se baseiam no volume e na integridade das conversas que acontecem no Twitter. Para volume, analisamos quantas pessoas estão Tweetando, Retweetando, respondendo e curtindo Tweets sobre esse tópico.

Queremos que os tópicos no Twitter reflitam conversas amplas e duradouras que as pessoas têm sobre eventos, pessoas e itens a debater. Por isso, usamos aprendizado de máquina para localizar Tweets relacionados dessas conversas. Isso pode significar que eles Tweetam muito sobre o tópico ou interagem muito com Tweets sobre o tópico. Daí em diante, encontramos os Tweets mais interessantes para essas pessoas, usando algoritmos, palavras-chave e sinais adicionais.

Nós implementamos diversas proteções para ajudar a manter as conversas incluídas em um tópico saudável e livre de abuso – inclusive trabalhando para não recomendar Tweets para um tópico se o engajamento é manipulado ou de baixa qualidade.

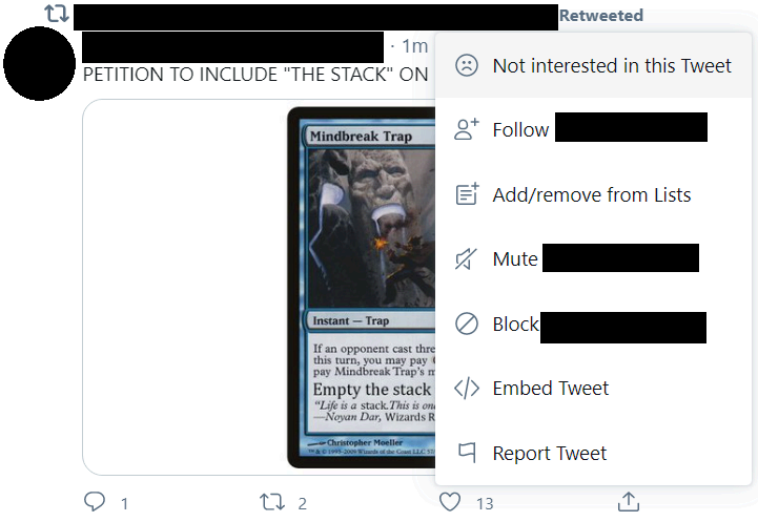
Fonte: Central de Ajuda do Twitter[15]

Para a segunda tarefa, é preciso encontrar uma publicação e selecionar a opção “Não tenho interesse neste tweet”, como mostra a Figura 10.

[14] Disponível em: https://help.twitter.com/pt/using-twitter/twitter-trending-faqs. Acesso em 6 de dez. de 2020

[15] Disponível em: https://help.twitter.com/pt/using-twitter/follow-and-unfollow-topics. Acesso em 6 de dez. de 2020

Figura 10 - Um menu com a opção de marcar desinteresse na publicação



Fonte: Tweet de usuário[16]

Analisando a distribuição dos signos nas interfaces para ambas tarefas, percebe-se que os signos estáticos e metalinguísticos são predominantes em relação aos signos dinâmicos. Na Central de Ajuda, há muitos links para artigos já na página inicial, como uma tentativa de dispor imediatamente artigos que podem ser de interesse do usuário. Entretanto, não se sabe se esses artigos são os mais visitados ou buscados, ou se os projetistas do site gostariam que os usuários vissem ao entrar. A barra de busca também permite que você busque artigos por todo o site, caso não encontre ele imediatamente na página inicial.

Dentro dos artigos em si, entende-se que o Twitter utiliza algoritmos para determinar que tópicos ou assuntos do momento são interessantes para os usuários. Embora seja um ponto positivo ter essa explicação em sua plataforma, não é possível encontrar uma forma de entender objetivamente o algoritmo em si. Como muitas outras redes sociais, o Twitter não revela como seu algoritmo funciona, o que não é a melhor maneira de garantir boa explicabilidade, mas como consta no estudo de Fjeld et al. (2020), é difícil ser transparente e privado, simultaneamente. Nesse mesmo estudo, é sugerido que os efeitos do algoritmo, e não o seu código fonte, sejam transparentes para que seja possível avaliá-lo.

Dito isso, continua-se para a reconstrução das metamensagens dos signos, começando pelo metalinguístico, que é:

[16] Disponível em: https://twitter.com/PleasantKenobi/status/1335670797059452929. Acesso em 6 de dez. de 2020.

- **Quem é o usuário:**

  "Entendemos que você é um usuário que entende rapidamente os conteúdos, sem precisar de muita explicação."
- **O que ele deseja fazer:**

  "Você gostaria de saber mais por que certos tweets aparecem em sua linha de tempo. Você também quer entender onde fica a seção de ajuda do site e saber se é possível realizar buscas sobre tópicos de ajuda e ler o conteúdo dos artigos resultantes."
- **De que formas prefere fazer:**

  "*Você quer fazer isso de forma rápida, lendo artigos ou explicações na própria interface."*
- **Qual sistema foi construído:**

  "Nós, designers do Twitter, projetamos este sistema. Com ele você pode tirar dúvidas e buscar ajuda sobre funcionalidades da ferramenta ou assuntos relacionados a elas."
- **De que formas deve ser utilizado:**

  "Para tanto, você deve entrar na seção de ajuda e buscar pelos tópicos desejados. Além disso, caso demonstre desinteresse em alguma publicação, deve selecionar a opção correspondente no próprio tweet."
- **Para atingir quais objetivos:**

  "Com esse sistema, você será capaz de entender o motivo de uma publicação ser recomendada para o seu perfil e melhorar a sua experiência no site."

Em seguida, a metamensagem dos signos estáticos:

- **Quem é o usuário:**

  "Entendemos que você é um usuário que gosta de explorar muitos artigos com uma frequência alta."
- **O que ele deseja fazer:**

  "Você quer interagir com conteúdo publicado por outros usuários da rede social e ver quantos usuários também interagiram. Além disso, você quer ser capaz de demonstrar seu desinteresse para nós."

- **De que formas prefere fazer:**

"Você quer ler vários artigos relacionados em sequência, curtir, comentar e retuitar publicações de outros usuários e demonstrar desinteresse, tudo isso de forma simples e rápida, a qualquer momento."

- **Qual sistema foi construído:**

  "Nós, designers do Twitter *projetamos este sistema. Com ele você pode tirar dúvidas e buscar ajuda sobre funcionalidades da ferramenta ou assuntos relacionados a elas, podendo navegar por diversas páginas que contém explicações que julgamos serem suficientes para sanar todas as dúvidas que possa ter. Você também pode interagir com o tweet dos usuários e ver quem interagiu. E se você não se interessar pelo tweet, você pode marcar seu desinteresse selecionando a opção correta no submenu, mas caso tenha feito isso por acidente, também é possível desfazer a opção."*

- **De que formas deve ser utilizado:**

  "A informação sobre interações é disponibilizada de forma compacta dentro do próprio tweet, ao lado das opções para curtir, compartilhar e retuitar. Para marcar o desinteresse, você deve selecionar a opção em um submenu. E caso deseje entender as funcionalidades do Twitter, você deve entrar na seção de ajuda, que acreditamos não ser tão frequentemente acessada, então colocamos dentro de um submenu no menu principal do site. Alternativamente, você pode digitar um termo de interesse na barra de busca para, e selecionar um artigo de interesse relacionado ao termo relacionado à funcionalidade que deseja entender."

- **Para atingir quais objetivos:**

  "Assim, você pode interagir com publicações no site e ver quem interagiu também. Você consegue, com o sistema projetado, entender as funcionalidades do Twitter."

Por fim, a metamensagem dos signos dinâmicos se torna:

- **Quem é o usuário:**

  "Entendemos *que você é um usuário que gosta muito de dicas e animações sutis na interface, mas acha legal movimentações mais explícitas na página. Finalmente, você gosta de ter muitas opções disponíveis em mãos, mas nem sempre expostas diretamente."*

- **O que ele deseja fazer:**

  "Você quer ler diversos artigos em nossa seção de ajuda a qualquer momento até o fim; e interagir com o conteúdo postado pelos usuários de diversas maneiras, sempre que for de seu interesse."

- **De que formas prefere fazer:**

  "Você quer ler artigos sem interrupções, e interagir com publicações de maneira rápida, mas com animações para ajudar a identificar o sucesso ou falha de uma interação."

- **Qual sistema foi construído:**

  "Nós, designers do Twitter *projetamos este sistema. Com ele você ler artigos em uma interface com poucos detalhes que podem te distrair. Você também pode interagir com o tweet dos usuários e ver quem interagiu."*

- **De que formas deve ser utilizado:**

  "Para ler os artigos, você pode entrar na nossa seção de ajuda e selecionar um artigo em alguma subseção de interesse na página principal ou buscar o termo relacionado ao assunto que você deseja na barra de busca. Após isso, selecione um resultado para ler o artigo. Para interagir com uma publicação, você pode clicar nas opções de curtir, comentar ou compartilhar, assim como marcar seu desinteresse sobre alguma publicação selecionando a opção em um submenu no tweet."

- **Para atingir quais objetivos:**

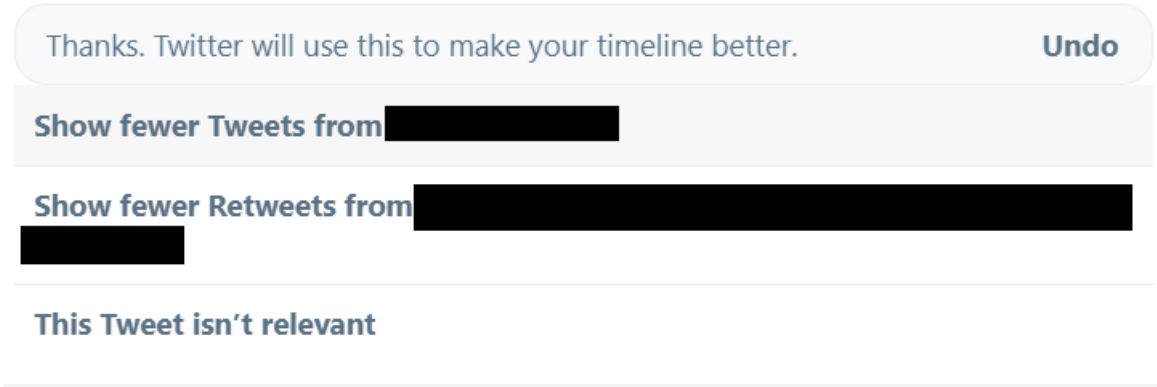
  "Com esse sistema, você é capaz de ler artigos sem interrupção, marcar seu desinteresse sobre uma publicação, o que irá melhorar a sua experiência de usuário. Nesse último caso, a publicação não será mais visível, e você terá a oportunidade de explicar o motivo de você se sentir assim, apesar de só terem disponíveis mensagens que consideramos relevantes para nós, então você terá que se contentar com mensagens genéricas."

Percebe-se que as estratégias de comunicação do Twitter quanto à transparência e explicabilidade consistem em artigos de ajuda superficiais e breves, para serem lidos rapidamente, e explicações na própria tela ao demonstrar desinteresse. Isso segue o padrão de comunicação dos projetistas visto na seção anterior, onde eles tentam garantir o acesso à informação, através de seções de ajuda, mas providenciam-nas imediatamente de forma superficial (no caso de demonstrar desinteresse). Entende-se que essas estratégias tentam dar mais controle ao usuário do site em relação à sua experiência, mas carecem de melhorias.

As três metamensagens mostram que os signos apresentam pouco contraste entre si e pouco redundantes, e no geral um signo não contradiz o outro. Por exemplo, apenas ao passar o mouse por cima de um *link* ele é realçado, mas estaticamente, o *link* não apresenta qualquer indicação de que pode ser interagido além de uma cor mais clara que outros textos para diferenciá-lo. Essa indicação ficou como responsabilidade do signo dinâmico, que escurece o texto ao passar o ponteiro do mouse por cima do *link*. Assim, nos signos estáticos, o uso de uma cor diferente para os *links* para informar que o esse texto é diferente dos outros condiz com o realce do signo dinâmico.

De qualquer maneira, é valido discutir que o Twitter não revelar seu algoritmo pode gerar impactos negativos relacionados à não-discriminação, algo também presente no estudo de Fjeld et al. (2020). Ao acumular dados suficientes, o Twitter pode utilizá-los para treinar seus algoritmos de IA e aprendizado de máquina e gerar vieses como efeito colateral, que irão crescer ainda mais caso não sejam corrigidos.

Figura 11 – Um menu com diversas opções para controlar a exibição de publicações



Fonte: Página Inicial do Twitter, após marcar desinteresse em um Tweet

Ao marcar desinteresse, o Twitter informa que essa ação será usada para melhorar a linha de tempo do usuário (Figura 11), mas fica a dúvida: Ao se mostrar desinteressado, **como as melhorias da linha do tempo dos usuários ocorrerão quando e com qual intensidade?** O que é outro ponto que prejudica o princípio de explicabilidade. No entanto, o fato de o conteúdo não ser mais exibido na tela após essa ação mostra bons valores de segurança por parte dos projetistas.

Ao interagir com uma publicação que está relacionada a um determinado tópico, não é informado imediatamente ao usuário se essa interação será suficiente para determinar os interesses e assuntos do momento ligados ao perfil que interagiu. É exibida apenas uma mensagem ou animação (no caso de curtidas) comunicando que a ação foi concluída com

sucesso. Isso pode ser considerado uma violação quanto à explicabilidade, já que o usuário sabe o efeito da ação, mas ele não é informado se o efeito ocorre ou não, como consta também em Fjeld et al. (2020). Por outro lado, o motivo dessa escolha de design pelos projetistas pode ter sido para evitar interrupções desnecessárias e não prejudicar a experiência do usuário, embora o impacto à explicabilidade se mantenha mesmo assim.

Finalmente, sobre os assuntos do momento especificamente, o artigo[17] na seção de ajuda explica que o usuário pode ver assuntos não personalizados de acordo com seus interesses, mas não informa como são determinados, mostrando mais uma fraqueza em relação a transparência e explicabilidade.

Concluindo, percebe-se que o Twitter se manteve moderadamente fechado e explicativo quando o assunto era explicar como os algoritmos funcionam objetivamente, ou avisar quando uma interação poderia impactar o algoritmo. Essas atitudes mostram dificuldades em promover transparência e explicabilidade por parte dos projetistas, e devem ser melhoradas para evitar consequências indesejadas aos usuários da rede.

4.3.3. Não-discriminação

Para análise do princípio da Não-Discriminação, o cenário elaborado foi: "*Alex gosta muito de redes sociais e se importa muito com acessibilidade e inclusão, já que tem de diferentes condições e culturas. Após ter uma experiência negativa com esses aspectos em outras redes, criou um perfil no* Twitter*, esperando que oferecesse opções melhores de acessibilidade de imagens e inclusão de gênero. As tarefas que deseja realizar são alterar o gênero de seu perfil e colocar descrições em imagens.*"

É possível identificar, portanto, duas tarefas nesse cenário: mudar o gênero do perfil e adicionar uma descrição a uma imagem. Essas são tarefas diretamente ligadas ao princípio de não-discriminação, já que envolvem indivíduos transgêneros e indivíduos com deficiência visual, que possuem dificuldades relacionadas à discriminação em ambientes que convivem.

Para publicar uma imagem, um usuário pode ou copiá-la e colá-la no campo de publicar um tweet, ou selecionar uma imagem armazenada no dispositivo. Após isso, o usuário deve selecionar a opção de adicionar uma descrição, onde aparecerá uma caixa de texto para que ele adicione uma descrição de até 1000 caracteres. Após salvar as alterações

[17] Disponível em: https://help.twitter.com/pt/using-twitter/twitter-trending-faqs. Acesso em 6 de dez. de 2020

por meio do botão “Save”, a descrição ficará disponível. A Figura 12 representa a interface dessa tarefa.

Figura 12 – Uma foto com campo para adicionar descrições a uma imagem para leitores de tela ou pessoas com baixa visão

Fonte: Twitter, ao editar uma imagem

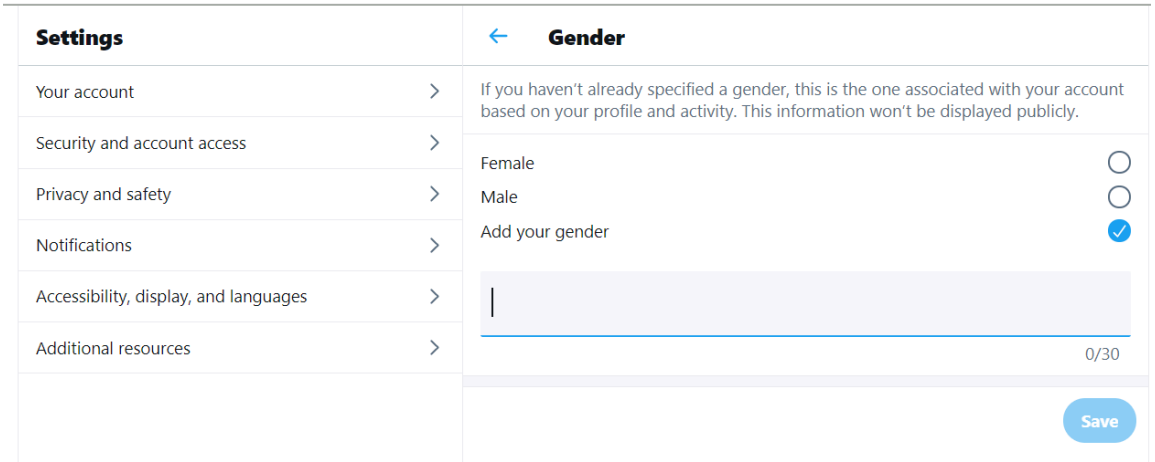
Inserir descrições a imagens é uma tarefa bem simples e fácil de encontrar, dado que a opção fica explícita inicialmente. Não há muitos signos metalinguísticos imediatamente visíveis que explicam o que são “descrições”, apenas um link para a Central de Ajuda. Por outro lado, há uma enorme quantidade de signos estáticos e dinâmicos, como mostrado na Figura 13. Ao selecionar uma imagem, diversos botões são habilitados na interface, o que pode prejudicar usuários com deficiência visual a identificar a opção correta. Após isso, porém, a interface é simplificada, contendo apenas a caixa de texto e alguns botões auxiliares, como pode ser visto na Figura 12.

Figura 13 – Um *card* com uma publicação antes de ser enviada para a rede

Fonte: Twitter, ao criar um tweet.[18]

Para a segunda tarefa, o usuário deve entrar no menu de configurações (como ocorreu em tarefas da seção de Privacidade e Segurança) e mudar o gênero na opção correspondente dentro da seção de "Informação da Conta" (veja Figura 14). Essa funcionalidade é acessível apenas após inserir corretamente a senha da conta.

Figura 14 – Botões de controle e caixa de texto para seleção de um gênero para a conta.



Fonte: Seus dados no Twitter[19]

[18] Disponível em: https://twitter.com/compose/tweet. Acesso em 6 de dez. de 2020.

[19] Disponível em: https://twitter.com/settings/your_twitter_data/gender. Acesso em 6 de dez. de 2020.

No menu de configuração do site estão presentes diversos signos estáticos, como ícones ao lado de títulos para representar uma seção, e setas direcionais para indicar navegação para outra seção; e metalinguísticos, como explicações abaixo do título dos menus e opções. Esses signos auxiliam a navegação do usuário pelos diversos menus, submenus e seções.

Seguindo para as metamensagens, iniciando com os signos metalinguísticos:

- **Quem é o usuário:**
  "Entendemos que você se importa com inclusão de gênero e acessibilidade."
- **O que ele deseja fazer:**
  "Você gostaria de saber o que e onde pode se expressar sobre o que está acontecendo. Além disso, entendemos que você quer saber de que maneiras é possível definir o gênero de seu perfil"
- **De que formas deseja fazer:**
  "Você quer fazer essas tarefas apenas quando não estiver percorrendo a linha de tempo"
- **Qual sistema foi construído:**
  "Visto isso, projetamos esse sistema onde há uma área onde você pode falar sobre assuntos de seu interesse e três opções de gênero para o seu perfil: masculino, feminino e adicionar o seu gênero."
- **De que formas deve ser utilizado:**
  "Para falar sobre assuntos de seu interesse, você deve clicar na área onde está escrito 'O que está acontecendo', digitar o que desejar (até 280 caracteres) e enviar a publicação.
  Para definir o gênero de seu perfil, você deve selecionar as uma das três opções, mas caso isso não seja feito, o gênero será definido automaticamente por nós, utilizando métodos que não revelamos."
- **Para atingir quais objetivos:**
  "Com esse sistema, você pode expressar o seu gênero livremente, e compartilhar suas opiniões sobre assuntos de seu interesse."

Já para os signos estáticos, temos:

- **Quem é o usuário:**
  "O nosso entendimento de você gosta de se expressar de muitas formas e se importa com inclusão de gênero e acessibilidade."

- **O que ele deseja fazer:**

  "Você também possui amigos e gostaria de marcá-los nas imagens que posta, e também quer editá-las ou adicionar uma descrição acessível para que leitores de tela consigam entender uma imagem. Também quer controlar quem pode responder aos seus tweets e publicá-los na rede. Finalmente, você gostaria de poder configurar algumas informações de seu perfil, incluindo gênero."
- **De que formas deseja fazer:**

  "Você pode realizar essas tarefas podem ser realizadas apenas ao editar um tweet que está sendo escrito."
- **Qual sistema foi construído:**

  "Construímos esse sistema onde é possível escrever e editar um tweet de diversas formas, como adicionar imagens, vídeos ou enquetes, além de escrever descrições para imagens contidas em um tweet."
- **De que formas deve ser utilizado:**

  "Para adicionar descrições a imagens, você pode clicar no botão 'Edit', habilitado enquanto um tweet está sendo escrito, e selecionar a opção 'Alt' para adicionar uma descrição. Similarmente, há botões para adicionar imagens, vídeos ou enquetes ao escrever o tweet.

  Sobre o gênero de seu perfil, é possível configurá-lo a partir de opções pré-definidas de gênero (masculino ou feminino), assim como adicionar um gênero que não seja um desses dois (de até 30 caracteres). Inclusive, essas informações são apenas acessíveis após digitar a sua senha, para protegê-las de pessoas não autorizadas."
- **Para atingir quais objetivos:**

  "Com esse sistema, esperamos que usuários incluam pessoas com pouca visão ou cegas em suas interações, assim como permitir que pessoas de gêneros que não masculino ou feminino possam informar o seu gênero."

Por fim, a curta metamensagem dos signos dinâmicos se torna:

- **Quem é o usuário:**

  "Nós do Twitter entendemos que você se importa com inclusão de gênero."

- **O que ele deseja fazer:**

  "Você quer ser capaz de definir o gênero de seu perfil."
- **De que formas deseja fazer:**

"Você gostaria de definir o gênero de maneira livre, e verificar se ele está correto."

- **Qual sistema foi construído:**

  "Entendido isso, construímos um sistema onde você pode definir o seu gênero e verificar se ele foi digitado corretamente, ou utilizar as outras opções pré-definidas por nós."

- **De que formas deve ser utilizado:**

  "Para definir o seu gênero, você pode digitá-lo, atentando para o limite de 30 caracteres, e ele aparecerá na tela para que possa verificar se foi digitado corretamente. Por outro lado, caso seu gênero seja masculino ou feminino, você também pode selecionar a opção correspondente dentre os botões radio."

- **Para atingir quais objetivos:**

  "Assim, você conseguirá incluir o seu gênero no seu perfil, embora de forma limitada."

Repara-se que o Twitter tem estratégias de comunicação um tanto confusas e até contraditórias quanto à não-discriminação. O uso de diversos botões ao publicar uma imagem ou texto é positivo e negativo, como será discutido adiante, e pode dificultar mais do que ajudar um usuário a inserir uma descrição na imagem. Isso dá a entender que o Twitter promove interação e comunicação entre os usuários da rede.

Apesar de prezarem pela não-discriminação, a grande quantidade de opções ao publicar algo pode discriminar e prejudicar usuários que ainda não se acostumaram com o site, impactando negativamente sua experiência. Além disso, usuários com deficiência visual podem ter dificuldade de identificar a opção de adicionar uma descrição dentre tantos botões, e nenhuma alternativa é fornecida para tal. Esse não é um fator diretamente associado à IA, mas pode gerar complicações que repercutem nos algoritmos, já que descrições são armazenadas como dados e podem contribuir para a definição de assuntos do momento, por exemplo.

A obrigatoriedade de selecionar um gênero através de botões de seleção única foge do padrão de comunicação que visava dar maior controle ao usuário, como através de caixas de seleção múltipla que o site utilizava até então, além de gerar diversos questionamentos não esclarecidos quanto à origem e finalidade desse tipo de dado.

Entretanto, pode-se perceber que o usuário consegue adicionar um gênero não especificado pelos projetistas, o que é um ponto muito forte em relação à não-discriminação, pois permite que o usuário determine seu gênero, ao invés de estar limitado a opções

pré-definidas. Assim, o Twitter retoma a estratégia de comunicação de dar controle ao usuário, promovendo a não-discriminação.

Especificamente sobre a interface de alterar o gênero, é interessante fazer algumas ressalvas. Há um signo metalinguístico na forma de explicação curta sobre as opções de gênero dentro do Twitter: se você não especificar um gênero, ele será definido pelo seu perfil e atividade (Figura 14). Esse é um problema tanto de não-discriminação quanto de transparência e explicabilidade. Fica a dúvida: **quais fatores são relevantes para determinar o gênero de um perfil?** Entende-se que isso é uma falha também pois o Twitter está gerando dados para motivos não conhecidos, já que esses dados não ficam visíveis publicamente para outros perfis acessarem.

Também não é claro o motivo do Twitter determinar o gênero de um perfil. **Para que fins esse dado é utilizado**? **Como ele é determinado**? **Onde posso encontrar mais informações sobre isso**? Nenhuma dessas perguntas é respondida, aparentemente, levando a uma questão fundamental: por que discriminar o gênero do usuário sem consentimento ou conhecimento dele, ainda que exista uma preocupação em promover a inclusão?

É possível ainda traçar um paralelo com privacidade e segurança, pois já que a informação de gênero compõe os dados da conta, algumas perguntas ficam em aberto: **os dados são compartilhados com terceiros?** Se sim, **quais as opções para anonimizar esse dado**, já que o perfil obrigatoriamente deve ter um gênero associado?

Assim, o Twitter novamente se mostrou falho nas suas estratégias de comunicação para promover valores éticos. Sobre a não-discriminação, o Twitter não explica muito sobre a finalidade e origem do gênero de um perfil. Por outro lado, promovem esse princípio ético através da livre escolha de gênero (embora obrigatório) e capacidade de adicionar descrições em imagens para melhorar a inclusão de usuários com deficiência visual.

4.3.4. Promoção de Valores Humanos

O cenário de inspeção do princípio ético Promoção de Valores Humanos elaborado é: "*Evelyn é uma usuária que prega muito a justiça social. Ela gosta de ver a promoção da cultura, inclusão e justiça nos conteúdos que consome. Por isso, segue muitos perfis que possuem representatividade de grupos minoritários e pessoas vulneráveis. Todavia, sabe que nem toda plataforma é perfeita, e após se deparar com algumas inconsistências, gostaria de entender como o* Twitter *lida com propagação de discursos ou ações de ódio.*"

Dessa vez há apenas uma tarefa a ser realizada: entender as políticas contra discurso de ódio na plataforma. Inicialmente, a inspeção é iniciada pela navegação por conteúdos contrastantes: um que favorece pessoas negras através de um ícone especial gerado automaticamente pela frase “#BlackLivesMatter” no Twitter, evidente na Figura 15; e outro que prejudica pessoas negras através da omissão de rostos negros em fotos devido ao algoritmo de reconhecimento facial em imagens utilizado pela rede social, representado nas Figuras 16 e 17. Em seguida, vemos um conteúdo falso que é denunciado pelo próprio Twitter.

Figura 15 - Um tweet com a frase "#BlackLivesMatter" e um ícone de punhos negros ao lado

Fonte: Tweet de David Lammy[20]

[20] Disponível em: https://twitter.com/DavidLammy/status/1317762298904301569. Acesso em: 6 de dez. de 2020

Figura 16 - Tweet onde parte de uma foto é omitida, marcada em vermelho

Fonte: Perfil de um usuário no Twitter[21]

Figura 17 - A foto que teve uma parte omitida (expandida)

[21] Disponível em: https://twitter.com/LittleMix/status/1242475345044860937. Acesso em 6. de dez de 2020

Fonte: Perfil de um usuário no Twitter[22]

Percebe-se que há uma maior utilização de signos estáticos e dinâmicos do que metalinguísticos nessas partes do site, perceptíveis através de imagens, que são interpretadas como dinâmicas devido ao redimensionamento e reconhecimento facial, e ícones representando estáticos, como ocorre com a frase-chave "#BlackLivesMatter" na Figura 15. Isso faz com que o princípio ético ocorra de forma simples, mas notável, o que é positivo já que não sobrecarrega o usuário com informações: não é qualquer frase-chave que possui ícones específicos, favorecendo o princípio ético de Promoção de Valores Humanos; e não é qualquer rosto que é omitido da foto, apenas o de uma pessoa negra, algo que viola o princípio ético de Não-Discriminação.

Quanto à ultima afirmação, percebe-se que esse é um problema tanto de não-discriminação quanto de promoção de valores humanos. Como consta no estudo de Fjeld et al. (2020), o princípio de Promoção de Valores Humanos envolve refletir nos algoritmos de IA os ideais culturais e sociais criados pelos humanos. Em outras palavras, o algoritmo deve representar os ideais de uma sociedade. No caso, são promovidos ideais de uma sociedade que não a negra, gerando discriminação.

Ademais, foi inspecionado também o fator de disseminação de notícias falsas. Uma notícia contida em uma publicação foi marcada como falsa a partir da anotação "*Multiple sources called this election differently*", mas havia poucos signos metalinguísticos para explicar o motivo e como isso ocorreu (se automaticamente ou manual), como mostram as Figuras 18 e 19.

[22] Disponível em: https://twitter.com/LittleMix/status/1242475345044860937. Acesso em 6. de dez de 2020

Figura 18 - Um tweet com um marcador que denuncia falsidade no resultado da eleição estadunidense de 2020

Fonte: Tweet do perfil de Donald Trump[23]

Figura 19 – Um tweet com uma notícia sobre o resultado da eleição estadunidense de 2020

Fonte: https://twitter.com/i/events/1318573265820921857. Acesso em: Acesso em 6 de dez. de 2020.

Diferente do que aconteceu nos exemplos anteriores, na publicação falsa e na notícia verdadeira, representadas pelas Figuras 18 e 19, respectivamente, há maior presença de elementos metalinguísticos. Apesar do link para o evento parecer ser um signo estático quando analisado por si só, ele atua como metalinguístico pois está vinculado ao conteúdo da

[23] Disponível em: https://twitter.com/realDonaldTrump/status/1330737141794676736. Acesso em 6 de dez. de 2020. Devido à suspensão do perfil, o tweet pode estar indisponível.

publicação. E no evento em si, há predominantemente signos estáticos e poucos signos dinâmicos, favorecendo um design simples e objetivo, mostrando que com poucos recursos é possível desmentir uma notícia.

Finalmente, a seção de ajuda contém um trecho sobre política de ódio (vide a Figura 20). Vemos que novamente os signos estáticos e metalinguísticos predominam. O site contém explicação bem definida e clara sobre como ele lida com propagação de ódio. Não há menção direta à IA, mas entende-se que discursos em larga escala serão mostrados para mais usuários frequentemente, devido a como o algoritmo do Twitter funciona, como visto na seção sobre Privacidade e Segurança.

Figura 20 – Trecho de artigo com explicação sobre as políticas de propagação de ódio no Twitter

Fonte: Central de Ajuda do Twitter[24]

A reconstrução da metamensagem dos signos metalinguísticos, portanto, nos dá:

- **Quem é o usuário:**
  "Você se importa muito com as 'fake news', respeito e diversidade."
- **O que ele deseja fazer:**
  "Você gostaria de poder identificar se as notícias são verdadeiras ou não, assim como saber nossas políticas contra propagação de ódio."
- **De que formas e por quê:**
  "A falsidade de uma notícia deve ser destacada de forma fácil de perceber, e devem existir explicações objetivas sobre as políticas contra propagação de ódio."

[24] Disponível em: https://help.twitter.com/pt/rules-and-policies/hateful-conduct-policy. Acesso em 6 de dez. de 2020

- **Qual sistema foi construído:**

  "Projetamos, portanto, um sistema onde notícias falsas são detectáveis e na seção de ajuda, é possível ver o nosso posicionamento contra propagação de ódio."

- **De que formas deve ser utilizado:**

  "Para detectar a falsidade de uma notícia, é exibida uma marcação no tweet, e ao clicar nela é exibida uma notícia real que contesta o que foi dito.

  Na seção de ajuda, é possível ler um artigo que explica objetivamente nossas políticas contra propagação de ódio no Twitter."

- **Para atingir quais objetivos:**

  "Com esse sistema, você consegue se alertar de e se prevenir contra notícias falsas. Você também pode entender o que caracteriza discursos de ódio e adaptar sua conduta no site para evitar punições, se necessário."

Sobre os signos estáticos, temos a seguinte metamensagem:

- **Quem é o usuário:**

  "Você que gosta de causas sociais e discriminação."

- **O que ele deseja fazer:**

  "Você ver fotos reduzidas para facilitar a visualização, e que hashtags populares sejam promovidas."

- **De que formas e por quê:**

  "Essas tarefas devem ser cumpridas de formas simples e rápidas."

- **Qual sistema foi construído:**

  "Visto isso, nós, projetistas do Twitter criamos esse sistema, onde hashtags populares relacionadas a causas sociais são acompanhas de um ícone especial e rostos de pele escura são omitidos de fotos através de algoritmos de reconhecimento facial."

- **De que formas deve ser utilizado:**

  "Ao ver com um tweet que contém uma hashtag que está popular no site, ela terá um ícone especial ao final da frase, automaticamente. As fotos também terão os rostos cortados automaticamente ao visualizá-las e para ver o rosto, você pode clicar na foto para ampliá-la."

- **Para atingir quais objetivos:**

  "Esperamos que você consiga ver a nossa promoção de causas sociais e a discriminação contra rostos de pele escura."

Para os signos dinâmicos, uma metamensagem é:

- **Quem é o usuário:**

  "Você entende que certas características devem ser discriminadas em relação a outras, e que notícias falsas devem ser combatidas com notícias verdadeiras."

- **O que ele deseja fazer:**

  "Você quer que as imagens publicadas no site sejam reduzidas e focadas em rostos de pele mais clara, assim como ver a notícia verdadeira que contesta uma falsa."

- **De que formas e por quê:**

  "Você não quer ter que controlar essas tarefas, delegando-as para o sistema."

- **Qual sistema foi construído:**

  "Assim, construímos esse sistema onde notícias falsas possuem um endereço que leva a uma notícia verdadeira que a contesta, e imagens são reduzidas automaticamente, caso sejam muito grandes."

- **De que formas deve ser utilizado:**

  "Para ver a notícia verdadeira, você deve clicar na marcação, sendo redirecionado para uma página que contém um resumo dela.

  E sobre reduzir imagens e focá-las em rostos brancos, para que sejam visíveis, você não precisa tomar ações: nós fazemos isso automaticamente utilizando nossos algoritmos."

- **Para atingir quais objetivos:**

  "Assim, você pode entender o que realmente ocorreu em um evento, além de não precisar ver rostos de pele escura em imagens no site."

Aqui, a estratégia de comunicação é feita através de signos ou explicações claros, simples e objetivos: são utilizados pequenos ícones para promover eventos sociais; redimensionamento de imagem que recorta partes não relevantes de imagens; artigos explicativos com tópicos objetivos e bem escolhidos; e notícias que desmentem conteúdo falso publicado. Embora essa estratégia promova, em certos casos, valores humanos, não há explicação do motivo de acontecerem, o que pode confundir um usuário. Essa é uma

diferença com a estratégia de comunicação dos projetistas até agora, que sempre procuram explicar o motivo de uma ocorrência.

Além disso, percebe-se que pela primeira vez o usuário não pode controlar nenhuma das evidências encontradas: não é possível denunciar notícias falsas deliberadamente; não é possível escolher ícones para frases-chave; e não é possível escolher qual parte da foto deve ser focada ao redimensionar, indo contra as estratégias de comunicação do Twitter até agora.

Por fim, embora eles tentem favorecer certos tipos de atitude através da denúncia de conteúdo falso e políticas contra disseminação de ódio, os projetistas implementaram um algoritmo de reconhecimento facial que faz exatamente o que eles não gostariam que fosse feito: prejudicar um grupo social protegido, como consta no artigo. Esse é um problema decorrente de viés em algoritmos de IA, que muitas vezes decorre do conjunto de dados utilizados para o desenvolvimento e treinamento deles. De fato, o estudo nota que se deve empregar a IA a benefício da sociedade, o que não aparenta ser o caso, já que prejudica uma etnia em certas ocasiões.

4.3.5. Considerações Finais sobre a Aplicação do MIS com Princípios Éticos

Após a aplicação do MIS, análise das estratégicas de comunicação e identificação das violações, entendemos que o Twitter apresenta estratégias de comunicação que buscam explicar e esclarecer o máximo possível para os usuários o motivo de um elemento ser exibido na interface, o que favorece os princípios éticos de Transparência e Explicabilidade.

Além disso, o Twitter possui estratégias de comunicação que favorecem a Privacidade e Segurança, permitindo que as pessoas usuárias configurem diversas opções de privacidade e segurança de suas contas, embora viole um pouco esses princípios a partir da falta de comunicação sobre o controle dos dados pessoais pelas pessoas usuárias.

Quanto à Não-Discriminação e Promoção de Valores Humanos, o Twitter é um tanto ambíguo nas estratégias de comunicação, favorecendo certos grupos (pessoas brancas e pessoas familiares com TICs) em detrimento de outros (pessoas negras e pessoas pouco familiares com TICs). Mesmo assim, o Twitter busca favorecer esses princípios éticos, apresentando políticas contra discursos de ódio e marcadores de notícias falsas. Para este último, no entanto, carece de explicações sobre os critérios para a marcação da notícia como falsa, prejudicando os princípios éticos de Transparência e Explicabilidade.

Entretanto, certos elementos, geralmente relacionados a algoritmos de IA ou dados utilizados ou gerados por eles, a explicação é superficial, deixando algumas perguntas em aberto. Isso foi notado, por exemplo, ao alterar o gênero de um perfil, onde o Twitter diz que esse gênero é determinado automaticamente caso não seja escolhido pela pessoa usuária, sendo considerado uma violação tanto aos princípios de Transparência e Explicabilidade quanto de Não-Discriminação.

Podemos ver, então, que os princípios éticos estão relacionados, e uma violação de um princípio ético, que ocorre por meio de uma estratégia de comunicação pouco favorável, pode prejudicar outro princípio ético.

4.4.Survey com usuários do Twitter

A pesquisa com os usuários do Twitter foi realizada através de um *survey*, com o **objetivo** de entender como os usuários lidam e percebem com os princípios éticos para IA (analisados com o MIS, veja seção 4.3) e também coletar fatos ocorridos com o uso do Twitter. Após a realização de um teste piloto, o *survey* foi distribuído de forma aberta pela Internet pela plataforma Google Forms (GOOGLE FORMULÁRIO, 2021) em busca de usuários regulares do Twitter, sem restrição de idade, localização, gênero e etc. O questionário apresentou e coletou o termo de consentimento dos participantes e 12 questões abertas e 23 questões fechadas. As perguntas foram agrupadas em 5 categorias:

- Demográficas: questões que indicam a representatividade da amostra, com dados, como por exemplo, local de moradia, grau de escolaridade e etc.;
- Sensíveis: questões que identificam, por exemplo, etnias, orientações sexuais e complementam nosso conhecimento sobre o perfil dos participantes;
- Experiência com Uso de Tecnologia: perguntas para identificar experiência no uso de tecnologias da informação e o perfil de uso do Twitter;
- Principais: questões que investigam opiniões e fatos sobre os princípios éticos; e
- Complementares: perguntas opcionais sobre o entendimento de ética dos usuários, o quanto discutem sobre os princípios éticos, o quão incluídos se sentem no Twitter e que conteúdo gostariam de ver no site.

O desenho do questionário levou em consideração alguns cuidados para diminuir os riscos na coleta dos dados: linguagem simples (sem termos muito técnicos); questões com apenas uma pergunta; questões sem duplo negativo, questões que investigam situações

recentes. Além disso, incluímos questões para coletar opiniões e fatos vividos pelos participantes.

No total, foram obtidas 165 respostas para o formulário, embora algumas perguntas abertas tenham sido ignoradas por alguns participantes.

Após a análise dos dados coletados nesta etapa realizamos uma etapa final para: a) verificar se os problemas éticos encontrados nas inspeções preliminar e semiótica são também de fato percebidos pelos usuários; b) descobrir outras violações dos princípios éticos para IA.

4.4.1. Análise do Perfil dos respondentes

As perguntas iniciais para identificação do perfil dos participantes incluíram as demográficas, de experiência com tecnologia e sensíveis. A distribuição na Internet, sem recrutamento prévio, buscou um perfil heterogêneo dos participantes para diminuir possíveis vieses na pesquisa.

Os participantes em sua maioria têm entre 18 e 30 anos (66,1% das respostas). Além disso, a maior parte das respostas foram dadas por pessoas com ensino superior, tanto graduação (50,9%) quanto pós-graduação (43,6%). Apenas 8 pessoas com o ensino médio e 1 participante que ainda não o completou.

Foi perguntado a etnia com que os participantes se identificam ou declaram. Muitos se declararam brancos (60%); 26,7% se identificam como pardo; 9,7% como negros; e 3 pessoas se autodeclaram como amarelas.

De acordo com as 153 respostas sobre a identidade de gênero, os participantes são em sua maioria mulheres (52,9%), seguido de homens (40,5%) e gêneros não-binários (6,6%). Ademais, 7 respondentes se consideram pessoas com deficiência, possuindo autismo, baixa visão, depressão, fibromialgia, surdez parcial, "*perna mais curta*" e *"uso de óculos"*. Sobre a orientação sexual dos participantes, aproximadamente 54,2% dos participantes que a informaram são heterossexuais, com o restante dividido entre homossexuais e bissexuais.

Por fim, nenhum participante é estranho à tecnologia, com 69,7% deles sendo muito familiares com a tecnologia.

Sobre o estado onde residem, obteve-se 163 respostas, das quais aproximadamente 65% correspondem à região Sudeste; 9,8% à região Sul; 8,6% da região Centro-Oeste; 8% à região Norte; e 8,6% à região Nordeste.

4.4.2. Análises dos Problemas éticos enfrentados durante o uso do Twitter

Esta seção apresenta os resultados das perguntas que investigam violações relacionadas à privacidade e segurança, transparência e explicabilidade, não-discriminação e promoção de valores humanos. Algumas perguntas tem resposta aberta para que o usuário possa se posicionar livremente, enquanto outras seguem a escala Likert utilizando em sua maioria escalas de 1 a 5 pontos. Assim, o menor valor corresponde a baixa frequência ou intensidade em relação a pergunta; e o maior valor, analogamente, corresponde a alta frequência ou intensidade. Para entender o posicionamento dos participantes, mapeou-se respostas comuns para as perguntas, gerando diferentes categorias de significado para a pesquisa.

De acordo com as respostas, pode-se perceber que apenas 10,3% dos usuários entendem bem quem usa os dados pessoais gerados ao interagirem no Twitter. Em outras palavras, apenas 16 pessoas sabem quem está utilizando as informações relacionadas a interesses pessoais. 61,2% dos participantes também responderam que veem anúncios relacionados ao conteúdo que interagem, com frequência moderada à grande. Isso significa que, ainda que eles não saibam quem usa seus dados, eles são utilizados mesmo assim, mostrando que há uma desinformação e uso não consentido dos dados por terceiros. Além disso, apenas 9,7% das pessoas pediram ao Twitter que seus dados fossem removidos do sistema, algo que favorece esse uso de dados.

O Twitter também configura as opções de privacidade de um perfil automaticamente ao criar um perfil, e essas configurações só mudam caso o usuário altere-as. Nesse caso, o compartilhamento de dados com outras entidades é habilitado por padrão. Nesta pesquisa, 57,6% dos respondentes nunca alteraram essa opção. Esse padrão é outro fator que facilita o compartilhamento de dados não consentido ou desconhecido, já que só é possível parar de compartilhar dados quando se descobre que os dados estão sendo compartilhados.

Os dados coletados dizem que 34,5% dos participantes não se sentem confiantes quanto ao conhecimento de quem usa seus dados, e 31,5% não sabem se os usam ou não. Apenas 5 pessoas (3%) estão convencidas de que não são acessíveis por ninguém além deles, embora apenas 2 destas tenham alterado as configurações de privacidade padrões do Twitter.

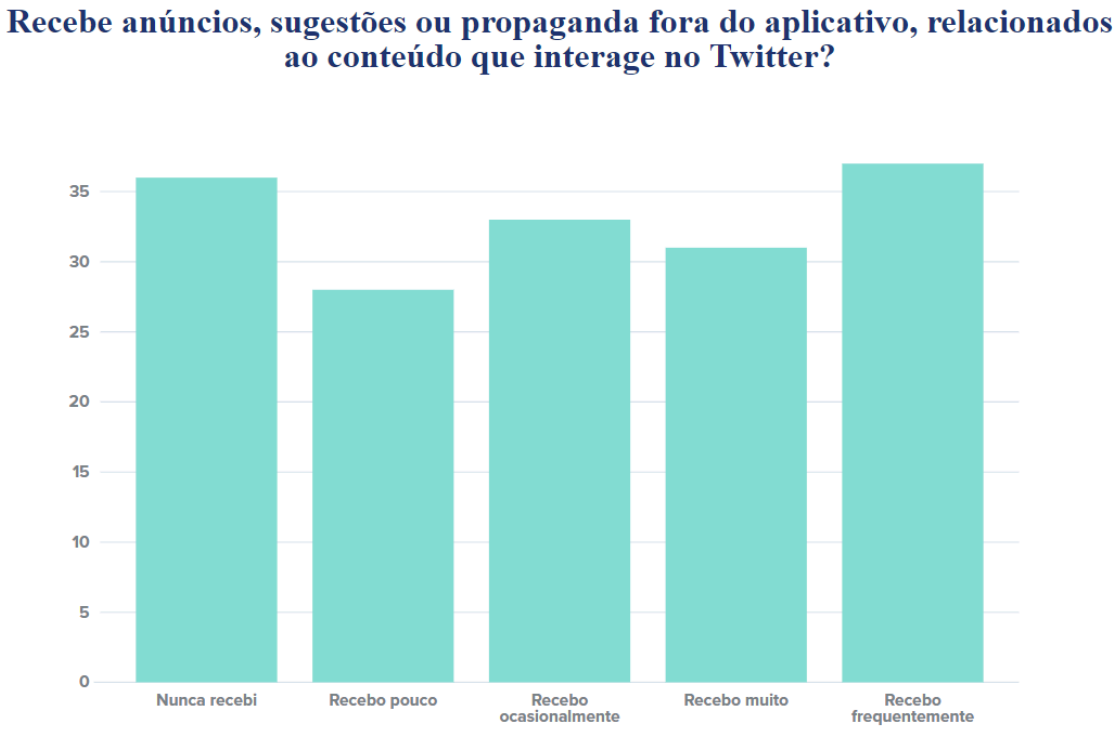
Quanto ao motivo de ver anúncios relacionados fora do Twitter, muitos participantes responderam que era devido a um “algoritmo”, “venda e compartilhamento de dados”, “cookies” ou “rastreamento de dados”. Todavia, usuários que responderam que não costumam publicar conteúdos no site tendem a ver menos anúncios e propagandas relacionados ao perfil da conta.

A quantidade de seguidores, pessoas que seguem ou frequência de uso do Twitter não são fatores principais e determinantes na quantidade de anúncios exibida, já que pessoas nas mesmas faixas de valores desses parâmetros apresentam variações extremas na quantidade de anúncios vista. Pessoas que utilizam pouco o Twitter tendem a ver um pouco menos de anúncios do que as pessoas que utilizam muito. Em média, as pessoas que utilizam o Twitter muito ou pouco veem anúncios com a mesma frequência, aproximadamente. Para estes dois extremos, em torno de 25% das pessoas veem anúncios com muita frequência e no geral, as pessoas tendem a ver com frequência intermediária.

Vemos, portanto, alguns problemas relacionados à **privacidade** dos usuários. O fato do Twitter não oferecer por padrão a privacidade dos dados facilita que eles sejam vendidos para outras organizações ou indivíduos. Essas organizações, então, usam os dados dos usuários sem o consentimento deles, e às vezes sem conhecimento também, o que pode comprometer a segurança dos usuários, pois não se sabe se os dados estão seguros com outras entidades.

Resumindo, não importa se você usa muito ou pouco o Twitter, os dados serão compartilhados independentemente. De fato, como mostra a Figura 21, as pessoas recebem quantidades diversas, mas próximas, de propaganda ou anúncios.

Figura 21 - Gráfico de barra com a frequência de recebimento de propagandas fora do Twitter pelas pessoas usuárias



Embora exista a LGPD, que visa proteger usuários do uso indevido de seus dados, apenas 43,9% das pessoas entendem bem o seu funcionamento para redes sociais e 59,4% nunca leu alguma explicação do Twitter sobre privacidade e proteção de dados. Isto pode ser usado a favor da venda de dados, já que o usuário não conhece seus direitos (LAUFER & WOLFE, 1977).

Investigando mais profundamente, das 64 pessoas que entendem pouco como funciona a LGPD para redes sociais, 51 nunca leram uma explicação do aplicativo sobre o assunto (quase 80%), enquanto das 101 que entendem bem, 53,4% já leram uma explicação do aplicativo sobre privacidade e proteção de dados pessoais.

Mesmo assim, apenas 7 respondentes revelaram que sofreram danos significativos devido à exposição de informações no site. Percebe-se que, pelo menos dentro desta rede social, a exposição e compartilhamento de dados não causam muitos danos perceptíveis. Isso pode ser devido ao Twitter manter dados relacionados a uma conta invisível para perfis externos, sendo então uma boa prática de segurança.

Ainda discutindo sobre dados, vamos analisar agora a utilização deles por algoritmos. No Twitter, sabe-se que os dados gerados pela interação no site são usados para mostrar conteúdos relacionados aos interesses do perfil. Entretanto, 82,4% das pessoas já viram algum conteúdo não relacionado aos interesses pessoais, ou que divergia das opiniões do usuário.

Mesmo com as explicações sobre o funcionamento do algoritmo, o Twitter parece não manter um sistema de recomendação razoável/adequado, visto que a maioria dos usuários já teve seus interesses ignorados em algum momento. Um possível motivo seria a intenção do Twitter de mostrar ao usuário conteúdos novos para ele consumir, ou diferentes opiniões para gerar uma discussão heterogênea. Entretanto, isso configura potenciais violações tanto relacionadas a explicabilidade e transparência dos algoritmos de recomendação, quanto a segurança do usuário, pois expor pessoas a opiniões muito contrárias pode causar danos a elas (ESTEBAN & SCHNEIDER, 2008), até mesmo por meio de discriminações e violações de direitos humanos.

Inclusive, sobre discriminação e valores humanos, 67,3% dos respondentes encontraram vieses na plataforma. Alguns respondentes alegaram terem visto "i*dentificação de rostos priorizar brancos*", "*fake news e hashtags relacionadas"* e "*discursos de ódio não punidos"*. Dentro desse contexto há, portanto, exemplos de **violação de graves problemas de discriminação e não promoção dos direitos humanos**.

Como visto durante o MIS (veja a seção 4.3), a priorização de rostos brancos em imagens também foi percebida pelos usuários do Twitter que responderam a pesquisa. A

presença de *fake news* foi verificada durante o MIS, embora nas respostas dos usuários elas não se mostrem punidas. Em específico, uma pessoa comentou que percebeu "*opiniões de governantes que não possuíam marcação de fake news*". Então, mesmo com a funcionalidade de detectar notícias falsas, parece que nem sempre essas notícias são devidamente marcadas. É visível, portanto, **a fuga ao princípio de não-discriminação nesse cenário**.

Sobre a promoção de valores humanos, o Twitter apresentou muitos problemas de acordo com o questionário. Diversas respostas revelaram que o Twitter não pune discursos de ódio na plataforma. Uma pessoa respondeu que já viu "Twitter *suspendendo contas de esquerda enquanto existem várias pessoas e robôs que espalham discurso de ódio e nada é feito*". "*Já vi vários perfis de mulheres lésbicas sendo bloqueados/inativados por falarem sobre sua vivência e opiniões.*", relata outra pessoa. Uma das respostas contesta o racismo na plataforma, dizendo: "[...] *não recebo indicações ou notificações de conteúdos que sejam não brancos, principalmente quando se trata de conteúdo LGBTQ+. Preciso pedir notificação ou 'lutar contra o algoritmo'*".

Entretanto, um respondente indicou que algumas denúncias são de fato aceitas: "F*oram algumas, inclusive denunciadas e que não foram aceitas. Várias denúncias foram aceitas. Muitas relacionadas à pedofilia, outras a questões de preconceito racial, de gênero*". Mesmo assim, a maioria das respostas relata que os discursos de ódio são constantemente ignorados, mesmo com as políticas contra discursos de ódio elaboradas pelos projetistas do Twitter. Isso é praticamente oposto ao que foi visto durante o MIS, quando se determinou que o Twitter emprega boas práticas de promoção de valores humanos. Esse contraste mostra que, potencialmente, sobre políticas elaboradas pelo Twitter, estudos com usuários são mais confiáveis quanto a eficácia das medidas elaboradas do que o MIS.

Dado tudo isso, as pessoas se sentem razoavelmente incluídas no Twitter, como mostra o gráfico na Figura 22, e o Twitter aparentemente mostra conteúdos de grupos minoritários para seus usuários além do que eles já consomem, de acordo com o gráfico da Figura 23.

Figura 22 - Gráfico de barras sobre sentimento de inclusão das pessoas usuárias no Twitter

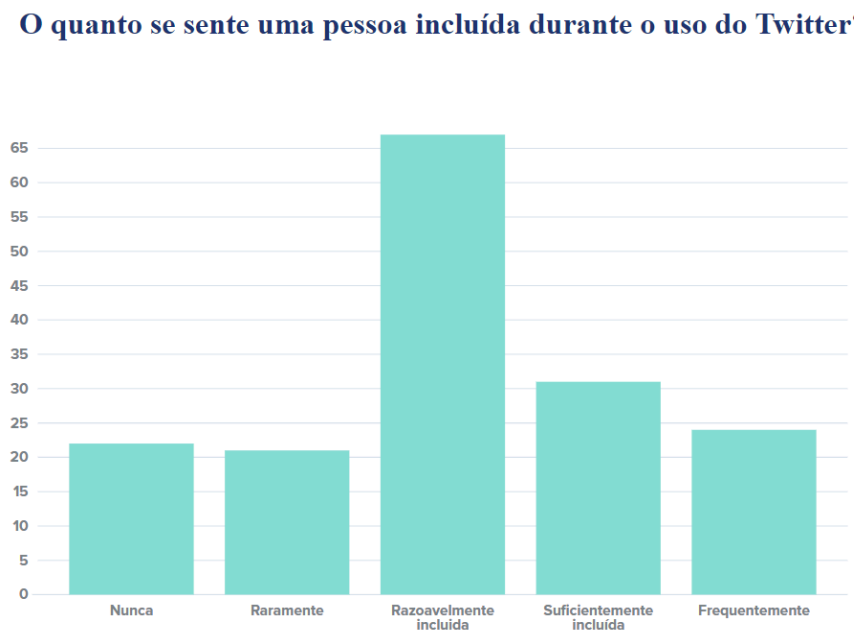
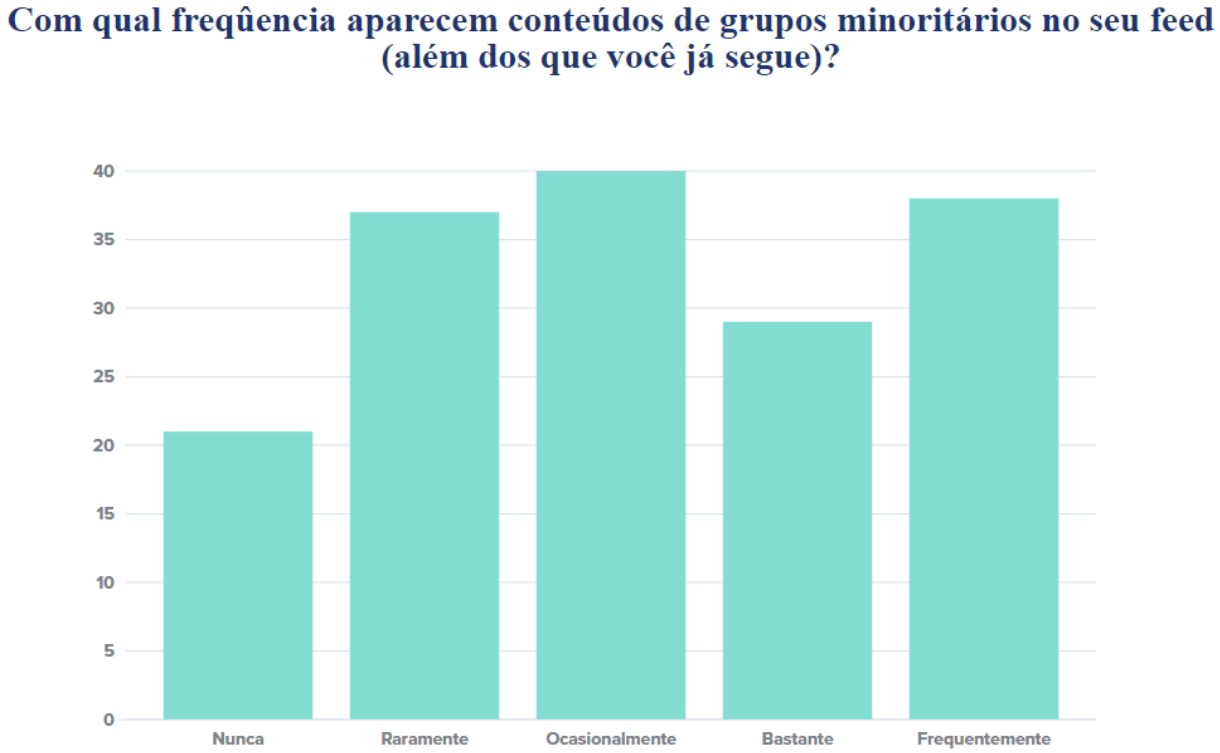


Figura 23 - Gráfico de barras sobre a frequência de consumo de conteúdo de minorias no Twitter pelas pessoas usuárias



Entretanto, os motivos que levam as pessoas a se sentirem pouco incluídas estão relacionados, dentre outros motivos, a discursos de ódio e falta de segurança na plataforma. Alguns usuários relataram que em algum momento se depararam com conteúdo ofensivo no

site, levando ao questionamento do funcionamento dos algoritmos de recomendação do Twitter. Se o que ele deveria mostrar está relacionado aos interesses do usuário, como ele mostrou algo que ofendeu um usuário, causando danos? Em outro ponto, por que as políticas contra discursos de ódio não puniram tais conteúdos, apesar de estarem definidas e abertamente acessíveis para os usuários?

Percebe-se então que a falta de promoção de valores humanos por parte do Twitter impacta a segurança dos usuários da rede. 29,7% dos respondentes já foram ofendidos após interagir abertamente no Twitter. Ainda que em quantidade menor do que os que não foram, é importante considerar que essas ocorrências englobam mais de um princípio ético, interagindo com segurança, transparência, não-discriminação e promoção de valores humanos.

Em suma, pôde-se identificar contrastes em relação ao MIS. Também foram encontrados mais problemas, descobertos através de relatos de usuários, que se mostraram fundamentais tanto para os contrastes quanto validação e extensão dos estudos feitos previamente.

4.4.3. Análise do Entendimento de Ética dos participantes

As perguntas finais do questionário pertencem à categoria de “Complementares”. Com elas, buscamos entender como os usuários enxergam ética e o quanto discutem sobre o assunto. Foram coletadas 158 respostas.

No geral, a ética está frequentemente associada a termos como “respeito”, “valores morais”, “regras coletivas”, “conduta”, “limites”, “discernimento” e “certo e errado” (veja Figura 24). Uma pessoa respondeu que ética “são princípios que direcionam o comportamento humano para que possamos ter uma sociedade mais justa, igualitária, pensando o respeito aos outros, a sociedade como um todo e o meio ambiente.”

Figura 24 - Uma nuvem de palavras sobre o entendimento dos respondentes sobre ética

Percebe-se então que para os usuários (pelo menos os que responderam a esta etapa), ética é algo que tenta manter o respeito entre um grupo, definindo regras para atingir esse objetivo através de conceitos capazes de discernir o "certo" do "errado". A partir disso, é notável que esses usuários entendem que é possível tomar comportamentos que prejudicam outras pessoas.

Não vimos diferenças no entendimento sobre ética, em função do perfil dos participantes. Por exemplo, daqueles que se autodeclararam brancos ou negros, todos entendem ética da mesma forma, com relatos citando "respeito" e "bem comum". Um respondente inclusive diz que ética é "*A busca por uma imparcialidade, na medida do possível, e a transparência das falhas e vieses desse processo. Escolhas de vieses que favorecem os direitos humanos e respeito pela diversidade. Tomada de escolhas*."

Os participantes que se autodeclararam brancos ou negros também entendem ética dessa forma, com relatos citando "respeito" e "bem comum". Um respondente inclusive diz que ética é "A busca por uma imparcialidade, na medida do possível, e a transparência das falhas e vieses desse processo. Escolhas de vieses que favorecem os direitos humanos e respeito pela diversidade. Tomada de escolhas."

O reflexo disso se encontra nas respostas sobre o que gostariam de ver no Twitter, que incluía em algumas respostas "conteúdo sobre grupos minoritários", "conteúdo que ajude a combater preconceitos" e "combate às *fake news*". Entretanto, essas pessoas parecem discutir pouco sobre ética, com aproximadamente 46,9% das pessoas discutindo pouco frequentemente ou raramente, e 28,4% discutindo com ocasionalmente. Mesmo assim, 67,7% delas também encontraram vieses na plataforma.

Outro dado interessante é que das 87 pessoas que explicaram o seu entendimento sobre ética, 60% também encontraram vieses na plataforma e 64,3% já foram ofendidas durante o uso do Twitter. Outro dado interessante é que apenas 12 pessoas nesse mesmo conjunto veem discursos de ódio com pouca frequência ou raramente, e 41 delas repara nisso quase sempre. Além disso, apenas 14 dessas pessoas se sentem muito bem ou bem incluídas no Twitter.

No contexto de LGDP, dessas mesmas 87 pessoas, 30 entendem pouco ou muito pouco sobre seu funcionamento, 22 entendem razoavelmente, e 35 entendem bem ou muito bem, e apenas 33 leram uma explicação sobre o assunto. Isso mostra que embora problemas de discriminação e valores humanos sejam mais perceptíveis para elas, a promoção à privacidade e segurança é um assunto um pouco estranho para algumas dessas pessoas e o Twitter poderia apresentar soluções para amenizar esse problema.

4.5.Considerações finais sobre os estudos

De forma resumida, o MIS encontrou potenciais problemas de Privacidade e Segurança, Transparência e Explicabilidade, Não-Discriminação e Promoção de Valores Humanos no Twitter, nos cenários inspecionados.

Sobre Privacidade e Segurança, o Twitter não permite que os dados de um perfil sejam removidos e os armazena por tempo indeterminado. Isso significa que o Twitter pode comercializar os seus dados sem que você tenha noção de quando os dados são e quando serão invalidados ou deletados. Ademais, a pesquisa relatou que muitos usuários não sabem quem utiliza seus dados, e aproximadamente metade dos que entendem bem quem utiliza os dados pessoais fora do Twitter já leram algum informativo sobre LGPD disponibilizado pela plataforma. Assim, os problemas de Privacidade e Segurança do Twitter relatados na pesquisa com usuários ultrapassam o que foi encontrado pelo MIS.

Muitas vezes, o Twitter também buscou ser transparente sobre o contexto de um elemento ser exibido na interface, algo revelado pelo MIS. Por exemplo, nas configurações do perfil, cada seção ou item do menu é devidamente acompanhado por um texto explicativo e ícone ilustrador, para facilitar a identificação da opção que o usuário deseja escolher. Entretanto, as seções de ajuda que explicam o funcionamento dos algoritmos não possuem detalhes claros e objetivos, ao contrário destes outros elementos.

Em paralelo, a recomendação de publicações não relacionadas ao conteúdo consumido ou de interesse foi relatada pelos usuários na pesquisa. Assim, embora muitos usuários tenham visto publicações que não eram de interesse, o Twitter explica que o conteúdo

recomendado está de fato relacionado com esse conteúdo. Assim, embora a aplicação do MIS conclua que há um motivo para o conteúdo aparecer, não se sabe se o problema mencionado ocorreu devido a uma ação do usuário ou uma falha do algoritmo. O problema foi de fato encontrado em ambas inspeções e pelos usuários.

Sobre Não-Discriminação, os problemas encontrados em ambas as análises foram semelhantes. Foram relatados vieses de priorização de rostos brancos na plataforma, discursos de ódio não punidos e *fake news*, como visto nos resultados do MIS.

E ao contrário do que foi obtido pelo MIS sobre Promoção de Valores Humanos, os respondentes da pesquisa relataram problemas de inclusão, disseminação de discursos de ódio e vieses na recomendação de conteúdo, priorizando alguns grupos étnicos a outros. Na aplicação do MIS, concluímos que o Twitter emprega boas práticas de Promoção de Valores Humanos, o que não ocorreu na experiência de alguns usuários que participaram da pesquisa.

Assim pode-se considerar que pesquisa com usuários foi complementar ao Método de Inspeção Semiótica, encontrando problemas que vão além dos resultados da última e ainda confirmam outros.

Entretanto, em alguns casos, o Twitter favoreceu os princípios éticos, como é o caso da explicação de elementos da interface, disponibilização de opção para parar de compartilhar dados com entidades externas ao site e também invisibilidade de dados pessoais da conta para outros perfis que não o do usuário para garantir a privacidade e segurança dele. De fato, sobre este último tópico, pouquíssimos usuários relataram que sofreram danos pela exposição de informações na rede. Mesmo que a pesquisa com usuários tenha encontrado diversos problemas, o MIS também encontrou potenciais pontos positivos no Twitter.

Finalmente, as estratégias de comunicação do Twitter sobre os princípios éticos para IA, identificadas na aplicação do MIS, são:

- Privacidade e Segurança: explicar quais ações podem ser tomadas e dar controle ao usuário sobre os seus dados.
- Transparência e Explicabilidade: disponibilizar artigos de ajuda com explicações superficiais sobre o funcionamento da aplicação do Twitter.
- Não-Discriminação: favorecer a inclusão de usuários, através de funcionalidades como, por exemplo, descrições em imagens e livre escolha de gênero.
- Promoção de Valores Humanos: promover valores humanos através de signos e explicações simples, objetivos e claros, como ícones especiais para frases-chave e algoritmos de reconhecimento facial.

No Quadro 2 são apresentadas as estratégias de comunicação, os princípios éticos para IA e como essas estratégias são utilizadas para promover ou violar princípios éticos para IA:

Quadro 2 - Estratégias de comunicação dos projetistas do Twitter e os princípios éticos aos quais estão relacionadas

Estratégia de Comunicação	**O que é comunicado**	**Princípios Éticos relacionados**	**Promove Princípios Éticos?**	**Viola Princípios Éticos?**
Botões *radio* e caixas de seleção.	Garantir controle ao usuário sobre dados e o gênero de seu perfil.	Não-discriminação, Privacidade e Segurança, Transparência e Explicabilidade.	Sim, potencialmente, de acordo com o **MIS**, pois o usuário pode alinhar a sua identidade de gênero ao gênero do perfil; e também pode controlar o compartilhamento de dados com parceiros do Twitter.	Sim, potencialmente, de acordo com o **MIS**, pois embora permita escolher um gênero de acordo com a identidade do usuário, ainda é obrigatório escolher esse gênero, sem explicação do motivo para isso.
Explicações sobre as ações que um usuário pode tomar ou opções que pode escolher.	Melhorar a compreensão do usuário sobre o que pode e não controlar.	Privacidade e Segurança, Transparência e Explicabilidade, Não-discriminação.	Sim, potencialmente, de acordo com o **MIS**, pois o usuário é capaz de entender como pode controlar a sua privacidade e segurança e qual o	Não, de acordo com o **survey**, já que muitos respondentes informaram que não liam muito sobre os dados e não tiveram prejuízos decorrentes da falta de privacidade. Mas potencialmente

			efeito dessas ações.	sim, de acordo com o **MIS**, pois usuários mais experientes podem se sentir desconfortáveis com tantas explicações redundantes.
Campos de entrada de texto.	Promover a inclusão, livre expressão e não-discriminação.	Não-discriminação, Promoção de Valores Humanos, Privacidade e Segurança, Transparência e Explicabilidade.	Sim, potencialmente, de acordo com o **MIS**, pois permite que usuários cegos ou com pouca visão utilizem leitores de tela para entender imagens; e permite que indivíduos que não se identificam com os gêneros masculino ou feminino adicione um outro gênero.	Sim, potencialmente, de acordo com o **MIS**, pois embora um usuário possa adicionar um gênero qualquer, não se sabe se esse dado é compartilhado com terceiros ou não, podendo causar danos ao usuário, já que é um dado sensível para pessoas com gêneros diferentes de masculino e feminino.
Artigos com informações sobre funcionamento de algoritmos.	Facilitar o entendimento do usuário sobre as funcionalidades do Twitter	Transparência e Explicabilidade.	Sim, potencialmente, de acordo com o **MIS**, pois ajudam o usuário a entender em certo nível os motivos	Sim, potencialmente, de acordo com o **MIS** e com o **survey**, já que as explicações são breves e superficiais,

	que envolvem IA.		de um conteúdo ser recomendado para ele.	dificultando a compreensão de possíveis vieses ou divergências com o perfil do usuário.
Ícones especiais; algoritmos de reconhecimento facial.	Promover eventos especiais; destacar as partes mais relevantes de uma imagem.	Não-discriminação, Promoção de Valores Humanos, Privacidade e Segurança, Transparência e Explicabilidade	Sim, potencialmente, de acordo com o **MIS**, os ícones especiais permitem que causas sociais populares sejam promovidas através de frases-chave (*hashtags*) associadas a um ícone promocional.	Sim, de acordo com o **survey**, os algoritmos de reconhecimento facial possuem um viés que omite o rosto de pessoas com pele negra, enquanto destaca o rosto de pessoas com pele branca.

5. CONCLUSÕES

Os problemas éticos decorrentes da evolução das TICs têm sido cada vez mais notados e estudados na literatura. As redes sociais e *Big Data* estão no centro dessa evolução, com cada vez mais pessoas usando ativamente mídias sociais e consequente aumento do fluxo de dados que são aproveitados pelas empresas em seus algoritmos de IA embutidos nos produtos e serviços oferecidos. Na contramão desta evolução, surgem os potenciais problemas éticos, a exemplo de violações de privacidade e segurança, que devem ser corrigidos a fim de não causar danos aos usuários, nem prejudicar sua experiência de uso deles.

Neste trabalho, realizou-se uma avaliação na aplicação web da rede social Twitter com os objetivos de: a) entender a comunicabilidade dos tais princípios éticos para IA durante o seu uso; b) entender se as pessoas usuárias do Twitter, enfrentam ou percebem problemas éticos durante o uso cotidiano; c) entender a percepção das pessoas usuárias do Twitter sobre Ética.

Na primeira etapa do nosso estudo, realizamos uma inspeção preliminar, buscando evidências em sites de notícias na internet, artigos científicos e dentro da própria plataforma. Foram encontradas violações de princípios éticos para IA (FJELD et al., 2020): privacidade e segurança, transparência, promoção de valores humanos e não-discriminação.

Na segunda etapa do estudo aplicamos o Método de Inspeção Semiótica (de SOUZA et al., 2006; de SOUZA & LEITÃO, 2009). Com isso foi possível aprofundar a análise desses princípios e entender como se dão as estratégias de comunicação do Twitter. Foram encontrados problemas na comunicabilidade relacionados tanto aos algoritmos de IA utilizados pelo Twitter, quanto na própria interface construída pelos projetistas.

Na terceira etapa, realizamos uma pesquisa com usuários do Twitter por meio de um *survey* para entender como os usuários lidam e percebem com os princípios éticos para IA e também coletar fatos ocorridos com o uso do Twitter. De fato, os usuários relataram problemas parecidos aos encontrados, além de outras insatisfações como quantidade notável de *fake news* e discursos de ódio.

Finalmente, na quarta etapa analisamos os resultados da aplicação do MIS com os dados coletados por meio do *survey* para identificar convergências ou divergências entre os dois estudos. Os resultados indicam que, de fato, os usuários participantes do *survey* relataram problemas muito próximos aos resultados encontrados nas inspeções.

Os estudos também revelam que os usuários possuem um entendimento de ética baseado em regras coletivas, moral e respeito ao próximo, o que mostra a preocupação deles

em manter uma boa conduta na plataforma. Assim, eles se mostram mais propensos a encontrar problemas éticos, garantindo mais confiabilidade para as respostas do questionário e tornando a investigação por *surveys* uma possível alternativa ou complemento a avaliações feitas por inspeção.

Visto isso, entende-se que este trabalho encontrou que o Twitter tem pontos positivos na comunicabilidade de princípios éticos, como permitir que um usuário informe no seu perfil um gênero que não "masculino" ou "feminino", algo muito bom para promoção de valores humanos e não-discriminação. Entretanto, a análise dos algoritmos de IA e do comportamento da aplicação, feita com o Método de Inspeção Semiótica, além da pesquisa com usuários e contraste entre os resultados de ambas, encontraram diversos problemas éticos como os mencionados acima.

Por fim, foi possível descobrir o entendimento de ética dos participantes, mostrando que embora existam problemas éticos que podem ser descobertos através métodos de avaliação, também é possível encontrá-los através de um questionário com usuários.

As principais contribuições deste trabalho são: a) a identificação das estratégias de comunicação do Twitter; b) o entendimento do perfil ético de usuários do Twitter; c) o mapeamento de violações a princípios éticos no Twitter; d) a revelação de pontos de influência de IA no processo de violação de princípios éticos.

5.1.Limitações

Uma das limitações deste trabalho é a quantidade de princípios éticos analisados, quando comparada a quantidade de princípios éticos para IA no trabalho de Fjeld et al. (2020). Como visto durante os estudos, os princípios éticos estão relacionados entre si, e com essa redução, a quantidade de análises possíveis também é reduzida.

Além disso, o perfil dos participantes do questionário, embora diverso, não atingiu uma diversidade quantitativamente uniforme ou próxima a isso. Por exemplo, muitos usuários residiam na região Sudeste do Brasil, e todos os respondentes tinham familiaridade com a tecnologia. Isso interfere nas análises, pois essas pessoas podem não relatar problemas que outras pessoas perceberiam.

Ainda sobre os participantes, a pequena quantidade de respondentes, em contraste com o número total de usuários da plataforma, prejudica a pesquisa, e impossibilita a generalização dos resultados.

Outra limitação é a ausência de uma discussão sobre o tema a partir de legislações brasileiras e mundiais, o que impede uma visão legal dos problemas éticos encontrados.

Por fim, estudar uma rede social em constante mudança dificulta a formação de uma conclusão definitiva sobre os estudos. Mesmo após o final dos estudos realizados neste trabalho, outros problemas e ganhos éticos surgiram na plataforma e continuam surgindo com o tempo. Embora esta análise tenha sido feita no presente, os pontos analisados podem mudar em pouco tempo, modificando os resultados encontrados em um curto período.

5.2. Trabalhos Futuros

Considerando o estudo de Fjeld et al. (2020), trabalhos futuros podem realizar uma análise a partir de princípios diferentes dos vistos neste trabalho, como responsabilidade ou controle humano da tecnologia.

Como este trabalho analisou interface e algoritmos do Twitter, seria possível focar em apenas uma esfera para que se possa obter resultados mais profundos e específicos à área em questão (IA ou comunicabilidade). Além disso, em relação à IA, o trabalho não entra em méritos altamente técnicos da área, compreendendo-a mais superficialmente. Assim, uma análise mais técnica pode ser capaz de trazer à tona problemas mais objetivos do que os encontrados.

Outro possível trabalho futuro é analisar diferentes redes sociais buscando violações a esses ou outros princípios éticos. A literatura possui vasto material pesquisando problemas em redes sociais, e jornais e redes de notícia costumam relatar esses problemas também, facilitando o estudo dessas redes e seus serviços. Além disso, é possível analisar outras redes a partir dos problemas encontrados no Twitter, para avaliar se elas também os apresentam.

Outro trabalho futuro em potencial é o uso de legislações brasileiras e mundiais na discussão sobre os problemas éticos encontrados, principalmente sobre os problemas relacionados a IA, o que pode auxiliar a descoberta de mais problemas e reforçar os encontrados nos estudos.

Finalmente, um estudo do comportamento da aplicação também pode ser realizado utilizando outro método de inspeção, com foco, por exemplo, na usabilidade e questões cognitivas.

REFERÊNCIAS BILIOGRÁFICAS

AUERNHAMMER, J. Human-centered AI: The role of Human-centered Design Research in the development of AI. *In:* Design and Research Society International Conference, 24, 2020, Queensland. **Proceedings […]** *[S.l.]*, v. 5, n. 1, p. 1315-1333, 2020.

BEIGI, G.; LIU, H. **A Survey on Privacy in Social Media:** Identification, Mitigation, and Applications. *Transactions on Data Science*, New York: Association for Computing Machinery, v. 1, n. 1, p. 1-38, jan. de 2020.

BROWN, T. Design Thinking. *Harvard Business Review*, Brighton: Harvard Business Publishing, v.1, n. 1, p. 84-95, 2008.

BUCHANAN R. **Human dignity and human rights**: Thoughts on the principles of human-centered design. *Design Issues,* Massachusetts: MIT Press, v. 17, n. 3, p. 35-39, 1 de jul. de 2001.

BURR, C.; CRISTIANINI, N. Can Machines Read our Minds? **Minds and Machines**, [*S. l.*], v. 29, p. 461-494, 27 de mar. 2019.

CAREY-SIMOS, G. How Much Data Is Generated Every Minute On Social Media?. **Wersm**, 2015. Disponível em: https://wersm.com/how-much-data-is-generated-every-minute-on-social-media. Acesso em 2 de fev. de 2021

COECKELBERGH, M. Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. Science and Engineering Ethics, Berlin: Springer, v. 26, n. 4, p. 2051–2068, 24 de out. de 2019.

CUTLER, A.; PRIBIĆ, M.; HUMPHREY, L. Everyday Ethics for Artificial Intelligence. IBM Corporation, 2019.

DATAREPORTAL. **Global Social Media Stats**. 2021. Disponível em: https://datareportal.com/social-media-users. Acesso em 2 de fev. de 2021.

DEAN, B. How Many People Use Twitter in 2021? [New Twitter Stats]. **Backlinko**, 2021. Disponível em: https://backlinko.com/twitter-users. Acesso em: 27 de abr. de 2021.

DE SOUZA, C. et al. The Semiotic Inspection Method. *In:* BRAZILIAN SYMPOSIUM ON HUMAN FACTORS IN COMPUTING SYSTEMS, 7, 2006, Natal. **Proceedings [...]** New York: Association for Computing Machinery, p. 148-157, 2006.

DE SOUZA, C.; LEITÃO, C. Semiotic Engineering Methods for Scientific Research in HCI. *Synthesis Lectures on Human-Centered Informatics,* San Rafael: Morgan & Claypool, v. 2, n. 1, 2009.

ELSHERIEF, M. et al. **Peer to Peer Hate**: Hate Speech Instigators and Their Targets. *In:* INTERNATIONAL AAAI CONFERENCE ON WEB AND SOCIAL MEDIA, 12, 2018, Palo Alto. **Proceedings [...]** Palo Alto: The AAAI Press, v. 12, n. 1, p. 52-61, 2018. Disponível em: https://ojs.aaai.org/index.php/ICWSM/article/view/15022. Acesso em: 2 feb. 2021.

ENTEMAN, W. F. Stereotyping, Prejudice and Discrimination. *In:* LESTER, M. P.; ROSS, S. D. **Images That Injure**: Pictorial Stereotypes in the Media. Santa Barbara: Praeger Publishers, 2003, 336, cap. 2, p. 15-22.

ESTEBAN, J; SCHNEIDER, G.; **Polarization and Conflict**: Theoretical and Empirical Issues. *Journal of Peace Research*, *S.l.*, v. 45, n. 2, p. 131-141, mar. de 2008.

FERRER, X. et al. **Bias and Discrimination in AI**: a cross-disciplinary perspective. *IEEE Technology and Society Magazine*, New York: IEEE, v. 40, n. 2, 11 de ago. de 2020.

FJELD, J. et al. **Principled Artificial Intelligence**: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Massachusetts: Berkman Klein Center for Internet & Society, 2020.

FLORIDI, L. et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds & Machines, Berlin: Springer, v. 28, n. 4, p. 689–707, 26 de nov. de 2018.

GOOGLE FORMULÁRIO. **Crie lindos formulários**. Disponível em: https://www.google.com/intl/pt-BR/forms/about/. Acesso em: 20 de março de 2021.

GRITZALIS, D. et al. **History of information**: the case of privacy and security in social media. *In:* HISTORY OF INFORMATION CONFERENCE, 1, 2013, Athens. **Proceedings […]** Athens: Nomiki Bibliothiki, p. 283-310, 2014.

HARRIS, T. How Technology is Hijacking Your Mind — from a Magician and Google Design Ethicist. **medium.com**, 2016. Disponível em: https://medium.com/thrive-global/how-technology-hijacks-peoples-minds-from-a-magician-and-google-s-design-ethicist-56d62ef5edf3. Acesso em: 27 de fev. 2021.

JONES, S. **Doing social network ethics**: a critical, interdisciplinary approach. *Information Technology & People*, Bingley: Emerald Publishing, v. 30, n. 4, p. 910-926, 2017.

JORDAN, M. I.; MITCHELL, T. M. **Machine learning**: Trends, perspectives, and prospects. *Science*, v. 349, n. 6245, p. 255–260, 17 de jul. de 2015.

KAZAI, G.; YUSOF, I.; CLARKE, D. Personalised News and Blog Recommendations based on User Location, Facebook and Twitter User Profiling. *In*: PROCEEDINGS OF THE 39TH INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 39, 2016, Pisa. **Proceedings […]** New York: Association for Computing Machinery, v. 39, n.1, p. 1129–1132, 2016.

LAUFER, R. S; WOLFE, M. **Privacy as a Concept and a Social Issue**: A Multidimensional Developmental Theory. *Journal of Social Issues*, Hoboken: Wiley-Blackwell, v. 33 n. 3, p. 22–42, 14 de abr. de 2010.

LIGHT, B.; MCGRATH, K. **Ethics and social networking sites**: a disclosive analysis of Facebook. *Information Technology & People*, Bingley: Emerald Publishing, v. 23, n. 4, p. 290-311, 2010.

LGPD BRASIL. **O que muda com a nova lei de dados pessoais.** Disponível em: https://www.lgpdbrasil.com.br/o-que-muda-com-a-lei/. Acesso em: 23 de abr. de 2021.

LOMOTEY, R. K.; DETERS, R. Towards Knowledge Discovery in *Big Data*. *In:* IEEE INTERNATIONAL SYMPOSIUM ON SERVICE ORIENTED SYSTEM ENGINEERING**,** 8**,** 2014, Washington. **Proceedings […]** New York: IEEE, p. 181-191, 2014.

MARTINEZ, R. **Artificial intelligence**: Distinguishing between types & definitions. *Nevada Law Journal*, Nevada: Nevada Law Journal, v. 19, n. 3, p. 9-36, 28 de mai. de 2019.

METZEL, J. F. Information Technology and Human Rights. *Human Rights Quarterly*, Baltimore: The Johns Hopkins University Press, v. 18, n. 4, p. 705-746, nov. de 1996.

NADER, M. General Data Protection Regulation (GDPR): What you need to know to stay compliant. **CSO Online**, 2020. Disponível em: https://www.csoonline.com/article/3202771/general-data-protection-regulation-gdpr-requirements-deadlines-and-facts.html. Acesso em 24 de abr. de 2021.

NORMAN, D. A; DRAPER, S. W. **User Centered System Design**: New Perspectives on Human-Computer Interaction. Hillsdale: Lawrence Erlbaum Associates, Inc., 1986. 526 p.

OHLHORST, F. Social Media Risks Increasing in 2021. **Security Boulevard**, 2021. Disponível em: https://securityboulevard.com/2021/03/social-media-risks-increasing-in-2021/. Acesso em: 23 de abr. de 2021.

PAVLIUC, A. Online abuse of Kamala Harris shows that automated moderation doesn't stop trolls. **Scroll,** 2021. Disponível em: https://scroll.in/article/985137/online-abuse-of-kamala-harris-shows-that-automated-moderation-doesnt-stop-trolls. Acesso em 6 de mar. de 2021.

PEIRCE, C. S.; HOUSER, N. **The essential Peirce:** selected philosophical writings. Bloomington: Indiana University Press, 1998.

PEREIRA, R.; BARANAUSKAS, M. C. C.; LIU, K. An Essay on Human Values in HCI. *Journal on Interactive Systems*, Porto Alegre: Sociedade Brasileira de Computação, v. 9, n. 1, 2018. DOI: 10.5753/jis.2018.689. Disponível em: https://sol.sbc.org.br/journals/index.php/jis/article/view/689. Acesso em: 7 de maio de 2021.

PERGUNTAS frequentes sobre assuntos do momento no Twitter. Central de Ajuda. Disponível em: https://help.twitter.com/pt/using-twitter/twitter-trending-faqs. Acesso em: 19 de out de 2020.

PLAISANCE, P. L. **Transparency**: An Assessment of the Kantian Roots of a Key Element in Media Ethics Practice. *Journal of Mass Media Ethics*, Milton Park: Taylor & Francis, v. 22, n. 2-3, p. 187–207, 5 de dez. de 2007.

POLÍTICA contra propagação de ódio. Central de Ajuda. Disponível em: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy. Acesso em: 19 de out de 2020.

PRATES, R.O.; DE SOUZA, C.S.; BARBOSA, S.D.J. Communicability Evaluation Method for User Interfaces. *Interactions*, New York: Association for Computing Machinery, v. 7, n. 1, p. 33-38, jan. de 2000.

PRATT, M. **What is business intelligence?** Transforming data into business insights. **CIO**, 2019. Disponível em: https://www.cio.com/article/2439504/business-intelligence-definition-and-solutions.html. Acesso em 7 de mar. de 2021.

Redes sociais foram responsáveis por 21% das vendas em 2018, revela pesquisa**.** **E-commerce Brasil**, 2019. Disponível em: https://www.ecommercebrasil.com.br/noticias/redes-sociais-foram-responsaveis-21-vendas-2018-revela-pesquisa. Acesso em: 2 de fev. de 2021

REED, K. Reports on Twitter hack confirm existence of admin tool used for social media censorship. **World Socialist Web Site**, 2020. Disponível em: https://www.wsws.org/en/articles/2020/07/24/twit-j24.html. Acesso em: 24 de nov. de 2020.

RODRÍGUEZ-MAZAHUA, L. et al. **A general perspective of Big Data**: applications, tools, challenges and trends. New York: The Journal of Supercomputing, v. 8, n. 72, p. 3073–3113, 20 de ago. de 2015.

SAGIROGLU, S.; SINANC, D. Big Data: A review. *In:* INTERNATIONAL CONFERENCE ON COLLABORATION TECHNOLOGIES AND SYSTEMS*,* 9, 2013, San Diego. **Proceedings […]** New York: IEEE, p. 42-47, 2013.

SALGADO, L.C.C.; DE SOUZA, C.S.; LEITÃO, C.F. On the epistemic nature of cultural viewpoint metaphors. *In*: BRAZILIAN SYMPOSIUM ON HUMAN FACTORS IN

COMPUTER SYSTEMS, 10, 2011. **Proceedings [...]** Porto Alegre: Sociedade Brasileira de Computação, p. 23-32, 2011.

SALGADO, L.C.C.; LEITÃO, C.F.; DE SOUZA, C.S. **A Journey Through Cultures**: Metaphors for Guiding the Design of Cross-Cultural Interactive Systems. London: Springer, 2012, 141 p.

SELLEN, A. et al. Reflecting human values in the digital age. *Communications of the ACM*, New York: Association for Computing Machinery, v. 52, n. 3, p. 58-66, 2009.

SMITH, M. et al. Big Data privacy issues in public social media. *In:* IEEE INTERNATIONAL CONFERENCE ON DIGITAL ECOSYSTEMS AND TECHNOLOGIES, 6, 2012, Campione d'Italia. **Proceedings [...]** New York: IEEE, p. 1-6, 2012.

TURILLI, M.; FLORIDI, L. The ethics of information transparency. *Ethics and Information Technology*, Berlin: Springer, v. 11, n. 2, p. 105–112, 10 de mar. de 2009.

UNICEF. **Declaração Universal Dos Direitos Humanos**. Disponível em: https://www.unicef.org/brazil/declaracao-universal-dos-direitos-humanos. Acesso em: 31 de mar. de 2021.

VAN DEN HOVEN, J. et al. Privacy and Information Technology. *The Stanford Encyclopedia of Philosophy*. Stanford: The Metaphysics Research Lab, 30 de out. de 2019. Disponível em: https://plato.stanford.edu/entries/it-privacy/. Acesso em: 8 de fev. de 2021.

VINCENT, J. Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. **The Verge**, 2020**.** Disponível em: https://theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist. Acesso em: 24 de nov. de 2020.

WARREN, J. This is How the Instagram Algorithm Works in 2021. **Later**, 2021. Disponível em: https://later.com/blog/how-instagram-algorithm-works/ Acesso em: 2 de fev. de 2021.

WITTKOWER, D. E. Principles of anti-discriminatory design. *In:* 2016 IEEE INTERNATIONAL SYMPOSIUM ON ETHICS IN ENGINEERING, SCIENCE AND

TECHNOLOGY, 2, 2016, Vancouver. **Proceedings [...]** Vancouver: Curran Associates, Inc., v. 1, n. 2, p. 1-7, 2016.

WOOLLEY, S.; HOWARD, P. Introduction. *In:* WOOLLEY, S.; HOWARD, P. **Computational Propaganda**: Political Parties, Politicians, and Political Manipulation on Social Media. Oxford: Oxford University Press, 2019. 263 p., cap. 1, p. 3-15.

APÊNDICE A – ROTEIRO DE PERGUNTAS DA ETAPA DE PESQUISA COM USUÁRIOS DA REDE SOCIAL

Universidade Federal Fluminense - Instituto de Computação

QUESTIONÁRIO AOS USUÁRIOS DO TWITTER©

Seus dados pessoais:

1. Qual é a sua faixa etária?
2. Em qual unidade federativa você reside atualmente?
3. Com qual etnia você se identifica?
4. Qual é o seu nível de escolaridade?
5. Qual a sua familiaridade com tecnologia em geral?
6. Qual é o seu gênero?
7. Qual é a sua orientação sexual?
8. Você é uma pessoa com deficiência?
9. Qual deficiência possui?

Seu uso de redes sociais:

1. Com qual frequência usa o Twitter?
2. Quantos perfis você segue?
3. Quantos seguidores você possui?
4. Com qual tipo de conteúdo você costuma interagir?

Suas percepções sobre Privacidade e Segurança de dados no Twitter:

1. Você sabe quem está utilizando seus dados provenientes da utilização da rede social Twitter?
2. Já solicitou ao Twitter ou algum parceiro do Twitter que apagasse seus dados?
3. O quanto de prejuízo já teve devido à exposição de suas informações no aplicativo?

4. Já alterou as configurações de privacidade padrão do Twitter?
5. Entende como a Lei Geral de Proteção de Dados se insere no contexto de redes sociais?
6. Já leu alguma explicação do aplicativo sobre a privacidade e proteção dos seus dados?

Suas percepções sobre a Transparência do funcionamento do Twitter:

1. Alguma vez se deparou com um tweet que não fazia parte do conteúdo que você interage normalmente, ou que tinha opiniões divergentes à sua?
2. Recebe anúncios, sugestões ou propaganda fora do aplicativo, relacionados ao conteúdo que interage no Twitter?
3. Por que acha que isso acontece?

Suas percepções sobre Discriminação e Respeito à Diversidade:

1. Com qual frequência percebe discursos de ódio no Twitter?
2. Já sofreu algum tipo de ofensa (física ou verbal) por ter interagido abertamente no Twitter?
3. Já viu algum viés na plataforma que favorecia um grupo de pessoas em relação à outro?
4. Qual foi o viés encontrado?
5. Com qual freqûencia aparecem conteúdos de grupos minoritários no seu feed (além dos que você já segue)?
6. O quanto se sente uma pessoa incluída durante o uso do Twitter?
7. Assinale as opções que levam você a se sentir incluída ou incluído durante o uso do Twitter?
8. Assinale as opções que levam você a se sentir pouco ou nada incluída ou incluído durante o uso do Twitter?

Suas percepções sobre Inclusão e Ética:

1. O quanto se sente uma pessoa representada no conteúdo que aparece no seu feed?

2. Com qual frequência discute sobre assuntos relacionados à discriminação, inclusão, privacidade e segurança?
3. Qual conteúdo acha que deveria aparecer mais nos tópicos populares, feeds e hashtags?
4. O que entende por ética?

APÊNDICE B – TERMO DE CONSENTIMENTO UTILIZADO PARA A COLETA DE DADOS DA PESQUISA COM USUÁRIOS DO TWITTER©

Universidade Federal Fluminense - Instituto de Computação

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (ADAPTADO PARA PESQUISAS EM MEIOS VIRTUAIS)

Você está sendo convidado a participar da pesquisa para o projeto "Avaliação Ética do Twitter", de responsabilidade da Profa. Luciana Salgado e da aluna [REDACTED]. Este trabalho faz do Projeto Final de Aplicação da estudante [REDACTED], realizado no Instituto de Computação da UFF.

Todas as perguntas e todos os assuntos tratados no questionário levarão em conta sua experiência com o tema. O estudo terá seus dados coletados estatisticamente para os fins da pesquisa.

Durante todo o processo, se você se sentir desconfortável por qualquer motivo que seja, poderá não responder ou até mesmo cancelar a sua participação no mesmo momento sem nenhuma necessidade de explicação a este pesquisador ou a qualquer outra parte relacionada a esta pesquisa. Caso ocorra a interrupção dos questionários, todos os dados coletados até então serão completamente apagados e não entrarão para a análise dos dados da pesquisa.

Por fim, nós garantimos a sua confidencialidade e privacidade. Não iremos incluir, sob nenhuma hipótese ou circunstância, o nome ou outras informações pessoais, de qualquer participante ao qual tivemos contato para a realização deste trabalho, mesmo que o participante tenha recusado ou desistido da entrevista.

Como esta pesquisa é de participação voluntária, sem nenhum custo para o participante, seu consentimento poderá ser retirado a qualquer tempo, sem nenhuma espécie de prejuízo ou qualquer outra penalização. Além disso, esta pesquisa também não irá fornecer nenhum pagamento, em nenhuma forma, para aqueles que participarem do estudo. Para sanar qualquer dúvida referente aos procedimentos, riscos, benefícios e outros assuntos relacionados com a pesquisa, basta entrar em

contato com os pesquisadores responsáveis pela forma desejada presente no topo deste termo.