Ética na IA:

como desenvolver um sistema ético
para Inteligência Artificial

CAPPRA
INSTITUTE
FOR DATA
SCIENCE

CAPPRA INSITITUTE FOR DATA SCIENCE / 2021

A ética da Inteligência Artificial

Esse é um estudo sobre a implementação de Inteligência Artificial, mais especificamente sobre o comportamento desses sistemas quando desafiados sobre assobre as questões éticas que tangem a sociedade. Muitos trabalhos estão surgindo sobre esse tema, de uma forma geral centrados no debate conceitual, algo fundamental para o avanço desse tema prioritário na agenda de governos e negócios ao redor do mundo todo.

Aqui abordamos mais especificamente o sistema ético que precisa ser configurado para sustentação desses mecanismos autônomos, ou seja, a preocupação dos pesquisadores do Cappra Institute foi na identificação das fragilidades de tais normas quando transformadas em automação. Fizemos questão de encerrar esse estudo com um diagrama que ajuda na validação de fases para a concepção desses sistemas éticos, afinal, independente do tipo de inteligência que espera-se dessas máquinas, todas elas serão julgadas por seus comportamentos e desvios éticos quando ocorrerem.

Vou começar esse estudo com 5 perguntas que podem ajudar a guiar sua jornada de descobertas e vão provocar sua reflexão inicial sobre o tema:

1. Máquinas precisam ter comportamento ético?
2. Os valores morais de IA são herdados ou devem ser ensinados?
3. O viés pode ser considerado um desvio ético?
4. Quem são os responsáveis pelas falhas éticas dos sistemas de IA?
5. Você está preparado para julgar o comportamento de uma máquina?

Espero que tenha uma boa leitura, que a reflexão lhe provoque para a ação e que as ferramentas aqui disponíveis lhe ajudem neste caminho!

Ricardo Cappra
Fundador do Cappra Institute

Índice

01

o que é ética?

Você não pode ter ética na IA sem ter ética.

01

O que é ética?

Quando tomamos alguma decisão ou escolhemos agir de uma determinada forma, nos baseamos em fundamentos éticos para realizar essa escolha. Isso só é possível porque a ética nos permite pensar não só no que poderia acontecer, mas também no que consideramos correto acontecer, de modo que possamos interferir no curso dos eventos a partir de nossas ações. Isso acontece, por exemplo, quando estamos andando na rua e vemos uma pessoa cair, podemos parar e ajudar a pessoa a se levantar ou continuarmos andando, ou ainda, rirmos da pessoa. A escolha que tomamos baseia-se nos preceitos éticos que carregamos.

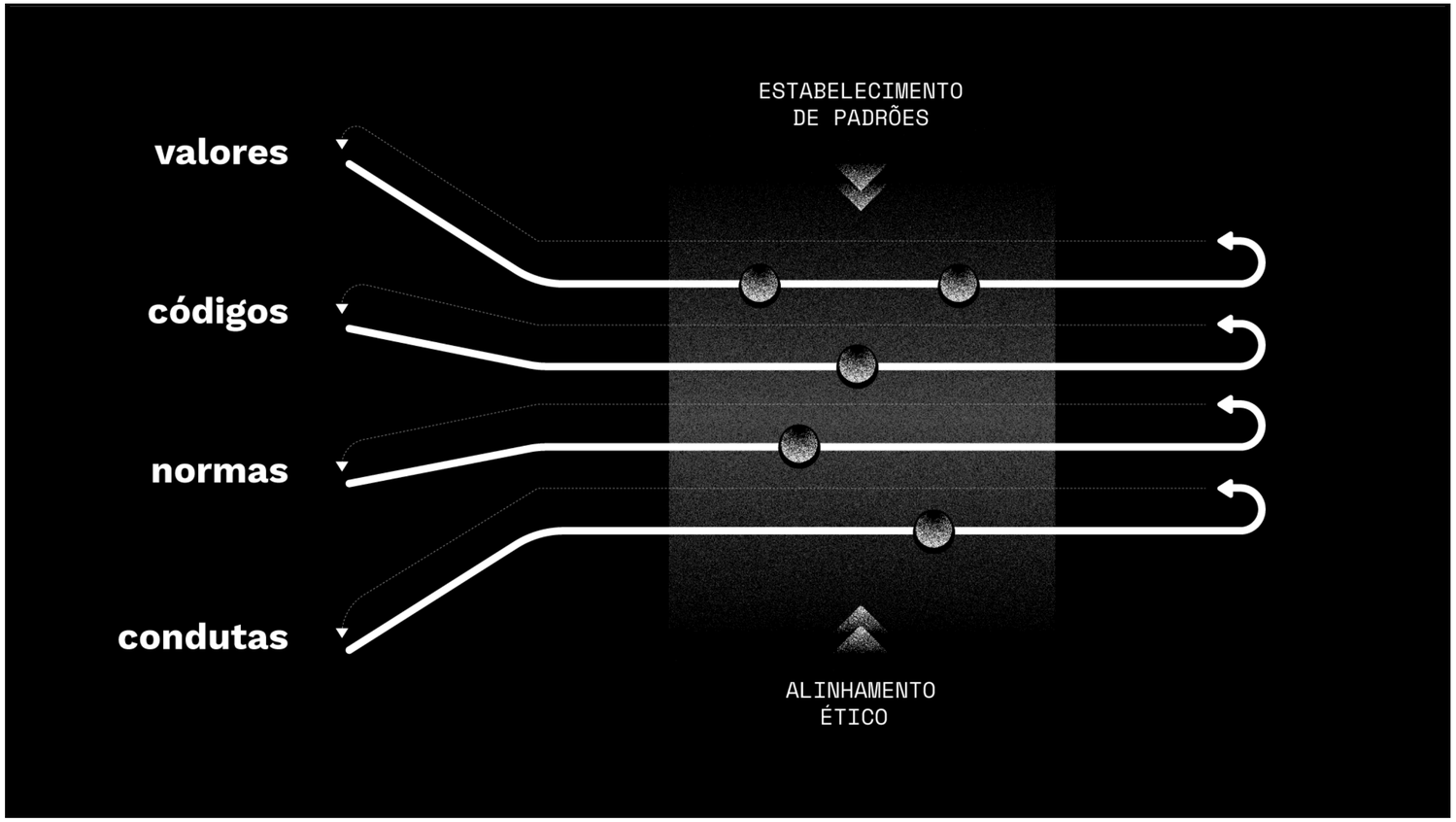
A ética, nesse sentido, apresenta-se enquanto uma reflexão sobre as nossas ações, mas também a adequação desse comportamento aos costumes vigentes locais. Os preceitos éticos possibilitam definir nossas ações com base nos padrões estabelecidos de certo e errado, independente das consequências. Sendo assim, ética é aquilo que consideramos como valores e ideais e que permeiam nossas escolhas e costumes.

Porém, a definição do que é ético não é absoluta, variando no tempo e no espaço. O que pode ser ético para um grupo de pessoas, não necessariamente será ético para um outro grupo. Da mesma forma, o que é considerado ético hoje, não necessariamente era considerado tempos atrás. Por exemplo, ter escravos no século XVI não era considerado antiético, pois era um comportamento aceitável para a época. Hoje esse comportamento seria considerado antiético por ir contra os valores de igualdade e liberdade estabelecidos na sociedade.

A definição de quais valores devem prevalecer está sempre em alteração. Sendo assim, os limites da ética estão sendo negociados e discutidos o tempo todo na sociedade. Por exemplo, muitas pessoas considerariam aceitável abrir mão de uma parcela de sua privacidade para aumentar a segurança, por outro lado, poderiam não aceitar essa mesma perda para receber melhores indicações de vendas ao comprar em um supermercado.

Os limites e os preceitos da ética são e devem ser sempre discutidos na sociedade, considerando as implicações que escolhas e ações podem ter não só na sociedade como um todo, mas também considerando grupos específicos. Entendermos que os valores e comportamentos atingem as pessoas de formas diferentes e que não há um sujeito universal, é fundamental para discutir e definir esses limites e implicações.

Os valores, códigos, normas e condutas estabelecidos como padrões a serem seguidos em relação às pessoas e às instituições na esfera do trabalho, formam a ética profissional, ou seja, a ética aplicada ao âmbito do trabalho. Nesse sentido, são estabelecidos padrões de conduta de relacionamento entre empregadores, empregados, clientes, prestadores de serviço e outros atores.



Quando colaboradores de uma empresa prejudicam colegas de trabalho para conseguir benefícios próprios, alguém em um cargo de liderança preenche uma vaga com um candidato que possui um relacionamento pessoal em detrimento de outros participantes do processo seletivo ou abusa moralmente de seus subordinados, temos exemplos de condutas antiéticas no ambiente de trabalho.

Falar sobre ética nas nossas relações pode parecer um tanto quanto complexo, mas quando inserimos máquinas e algoritmos a esta equação, podemos complicar ainda mais pois um algoritmo não pensa, e nem deve pensar, como a gente.

Moralmente óbvio para humanos. Não para máquinas.

Ação: robô à prova d'água vai ao correio postar uma carta.
Problema: o caminho é ao lado de um rio e uma criança de 5 anos cai dentro da água.
Para um humano, não há dúvida do que fazer: salvar a criança. Para a máquina também não há dúvida do que fazer, postar a carta no correio, que foi a atividade para a qual foi designada.
Agora, se o robô tivesse como escolher o que fazer dentre as duas ações (postar a carta ou salvar a criança), como poderia decidir? O robô precisa de regras para tomar decisões.
Digamos, então, que consideramos as seguintes métricas de sucesso:
Vida de uma criança = +1.000.000
Postar uma carta = +1
Finalmente o robô sabe como tomar a decisão e irá salvar a criança em detrimento de postar a carta imediatamente.
Mas, e se o robô estivesse dirigindo um caminhão com 1.000.001 cartas?
A lógica do robô seria:
Vida de uma criança = +1.000.000
Postar 1.000.001 cartas = +1 x 1.000.001 = 1.000.001.
Nesse caso, o robô iria deixar a criança se afogar.
Este exemplo mostra que esperar que máquinas tomem decisões eticamente corretas, depende da forma como as programamos. Suas decisões não são baseadas em ética e valores morais, mas sim em regras e dados.

Adaptado de "An Introduction to Ethics in Robotics and AI" (Christoph Bartneck et al, 2019).

02

ética tecnológica

Por que devemos falar de ética quando falamos em dados?

02

ética tecnológica

Atualmente a sociedade depende cada vez mais de informação, tornando possível chamá-la de sociedades data-driven, apoiada por tecnologia de informação. A combinação de acesso acelerado a grandes conjuntos de dados, abordagens algorítmicas melhoradas, avanços em hardware e otimização dos métodos de armazenamento, aumentou a necessidade de integrar cultura humana e digital. A produção contínua, automatizada e personalizada de estímulos e informações alcança um enorme número de pessoas aparelhadas e conectadas.

As pessoas recebem e transmitem sinais, atuando como peças-chave da engrenagem que mantém a circulação e o crescimento do volume de dados no mundo.

Há quinze anos atrás, as mídias sociais mal existiam. Hoje, muitos de nós começam o dia através delas. Ou seja, estamos completamente submersos em um mundo data-driven. Somos hoje 3.5 milhões de pessoas conectadas e, para exemplificar, geramos anualmente 200 Bi de tweets, 3.600 Bi de compartilhamentos no facebook e 360 Bi de horas de vídeos assistidos no youtube.

Todas as formas de conteúdo - midiáticos, profissionais, ou até mesmo conversas cotidianas - são influenciadas por algoritmos criados pelas redes sociais que disseminam informações. Dependendo da forma como esses algoritmos são criados, podem representar tanto oportunidades quanto ameaças para a sociedade e provocar rupturas e impactos em toda a malha social. As oportunidades podem ser trabalhadas através da ciência de dados e estão diretamente associadas a desafios éticos.

Mas por que falamos de ética quando falamos de dados?

A crescente quantidade e uso de dados, subproduto de processos do cotidiano em quase todas as áreas, tem facilitado a vida das pessoas por meio de uma nova forma de olhar para os dados. Dados de mobilidade se transformam em mapas, dados de clima se transformam em previsões do tempo, dados de busca se transformam em recomendações. Todas essas transformações são realizadas através de algoritmos e aumentam a compreensão das pessoas com relação ao mundo ao seu redor, consolidando o uso de técnicas de machine learning (ML) para a criação de soluções.

A crescente dependência de algoritmos para tomada de decisões e a redução gradual do envolvimento humano levantam preocupações com relação à justiça, responsabilidade, transparência e respeito aos direitos humanos na utilização dos dados de forma ética.

A ética no uso de dados se relaciona com as escolhas que fazemos no âmbito individual e coletivo, bem como as escolhas que as organizações fazem no âmbito empresarial. São elas que detém grande parte do poder de impactar a sociedade e sua ética está diretamente relacionada com o impacto que provocam.

Para exemplificar a ética das empresas (ou a falta dela), podemos pensar em exemplos como:

- Sua empresa compra dados de outras empresas que não deixem explícito de onde veio cada uma das informações comercializadas?
- Você já se cadastrou em uma promoção e, repentinamente, passou a receber ligações de vendas imobiliárias?
- É permitido utilizar dados como gênero, raça, religião, etnia nos algoritmos criados pela sua companhia?

A definição do que é ético ou não para cada organização é apenas o começo. Mesmo com definições claras, os dados podem carregar comportamentos não éticos e culturais, replicando a falta de ética em decisões baseadas em algoritmos e regras. Esse tipo de comportamento é o que chamamos de viés.

03

ética ou viés?

A relação entre falta de ética e viés

03

ética ou viés?

As pessoas que começam a se aventurar no caminho da Inteligência Artificial constantemente ouvem falar de viés. Também é constante a confusão entre viés e ética. Esses são conceitos que podem estar interligados, mas são diferentes.

Nas organizações, ser uma pessoa ética significa seguir os padrões pré-estabelecidos pela empresa, isto é, agir de forma coerente à governança estabelecida. Nesse sentido, um questionamento importante é se a ética das organizações atinge todas as áreas, incluindo as áreas responsáveis por analytics e modelos avançados, como IA. Assim, é necessário verificar não só se as práticas da empresa seguem preceitos éticos, mas também se o produto desenvolvido pela empresa, seja ele uma ferramenta ou um serviço, não fere questões éticas.

A ética relacionada ao uso de dados geralmente explicita o que pode ou não ser utilizado no desenvolvimento de modelos e regras. Muitas vezes os algoritmos criados buscam discriminar grupos (de pessoas, de objetos, de comportamentos), e, por esse motivo, algumas empresas evitam informações como raça, etnia, religião, sexo e gênero tentando se proteger de resultados antiéticos que possam gerar problemas de reputação. No entanto, utilizar os dados desses grupos minoritários na construção de inteligência pode reduzir a disparidade de ações direcionadas a estes públicos, além de evitar que seus dados se apaguem ou percam relevância, ou seja, a decisão de usar ou não esses dados depende da forma como a empresa os considera em suas análises.

Estamos vivendo um momento especial no qual é necessário encontrar o equilíbrio entre o avanço de técnicas sofisticadas e definições éticas. O avanço das técnicas de IA é condição obrigatória para que as empresas se mantenham competitivas. Por outro lado, as questões éticas tomam proporções maiores, não só por exigências legais, como a Lei Geral de Proteção de Dados (LGPD), mas também porque os consumidores se tornam cada vez mais exigentes com relação a questões sócio-político-ambientais, condições fundamentais para manter as pessoas interessadas na marca.

Mas, conforme dito anteriormente, a definição do que é ético é apenas o começo. Mesmo com definições claras, precisamos ainda lidar com o viés, que pode ser de vários tipos. O importante é compreendermos que o resultado de modelos e regras precisam ser avaliados minuciosamente, a fim de verificar se atendem aos padrões éticos ou não. Muitas vezes, mesmo não utilizando diretamente dados considerados antiéticos, as técnicas são capazes de diferenciar estes públicos e gerar resultados indesejados. Se isso acontecer, a falta de ética não identificada no desenvolvimento dos modelos e regras poderá se tornar um problema de proporções muito maiores ao aumentar a escala de abrangência. À medida que a escala de abrangência aumenta, a falta de ética aumentará proporcionalmente.

No ano de 2021, diversos documentos da Facebook Inc. foram vazados pela ex-engenheira de dados da empresa, Frances Haugen, em uma série de reportagens do Wall Street Journal. Uma das denúncias é que a empresa sabe que a rede social Instagram é danosa para meninas adolescentes pois bombardeia seus feeds com imagens de corpos perfeitos, mas não só escondeu o resultado dessa pesquisa como não fez nenhuma alteração na plataforma para mudar esse cenário.

Uma outra denúncia é que funcionários da empresa alertavam sobre contas de cartéis, pornografia e tráfico de pessoas, mas a empresa não fazia nada, ou muito pouco, para derrubar esses grupos.

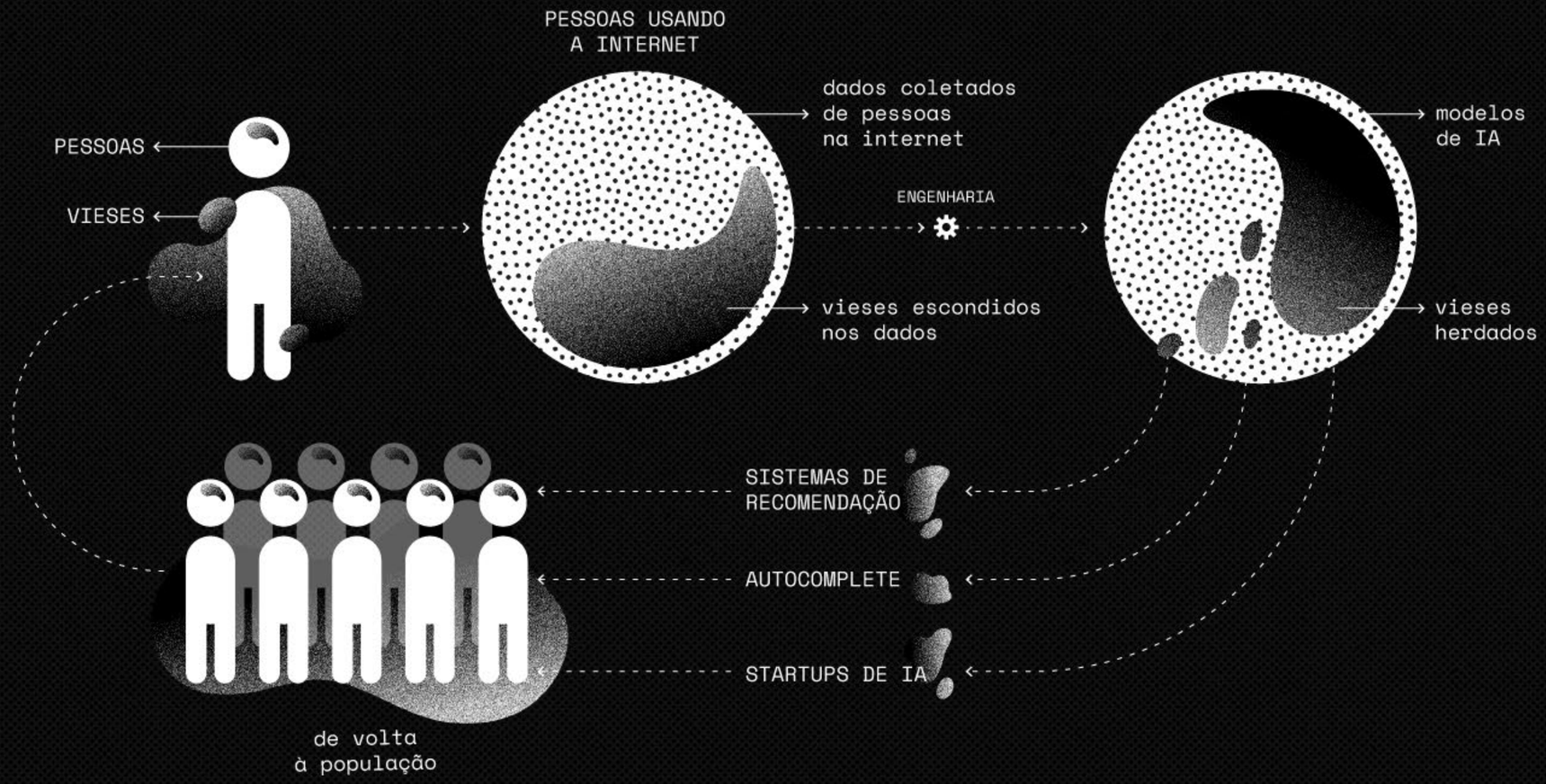
https://www.wsj.com/articles/the-facebook-files-11631713039

Um outro exemplo é o algoritmo de recomendação do YouTube, que possibilita a radicalização de seus usuários. Basta a pessoa acessar um vídeo que começa a receber recomendações de outros vídeos relacionados.

No caso de vídeos que são relacionados a ideias extremistas, o usuário vai se aprofundando cada vez mais nesses discursos.

https://theintercept.com/2019/01/09/youtube-direita/

Propagação de viés em dados



Ao utilizarem tecnologias de informação, as pessoas deixam rastros digitais em forma de dados, que são posteriormente transformados em modelos de Inteligência Artificial.

Mesmo alavancando o poder do Big Data, nenhum conjunto de dados conseguirá representar toda a população do planeta, fazendo com que esses conjuntos carreguem os vieses contidos no recorte selecionado, e consequentemente, os modelos também, sendo amplificados em seu retorno à população em um loop de feedback.

Entenda mais sobre esse conceito no capítulo a seguir.

04

design de algoritmos

Dados podem ser objetivos,
mas humanos não são

04

design de algoritmos

Normalmente, os algoritmos de ML operam aprendendo modelos a partir de dados existentes, e adaptando-os a dados novos. Como resultado, podem surgir problemas durante a coleta de dados, desenvolvimento de modelos e processos de implantação, que podem levar a diferentes consequências ao longo da sua operação. Este processo é longo e complexo, fundamentado no contexto histórico e movido por escolhas e normas humanas. Entender as implicações éticas de cada etapa do processo de geração e análise de dados pode revelar formas mais diretas e impactantes de prevenir ou lidar com as consequências prejudiciais nas atividades dos algoritmos. Além disso, é importante reconhecer que nem todos os problemas devem ser relacionados apenas aos dados, mas a todo o processo de criação.

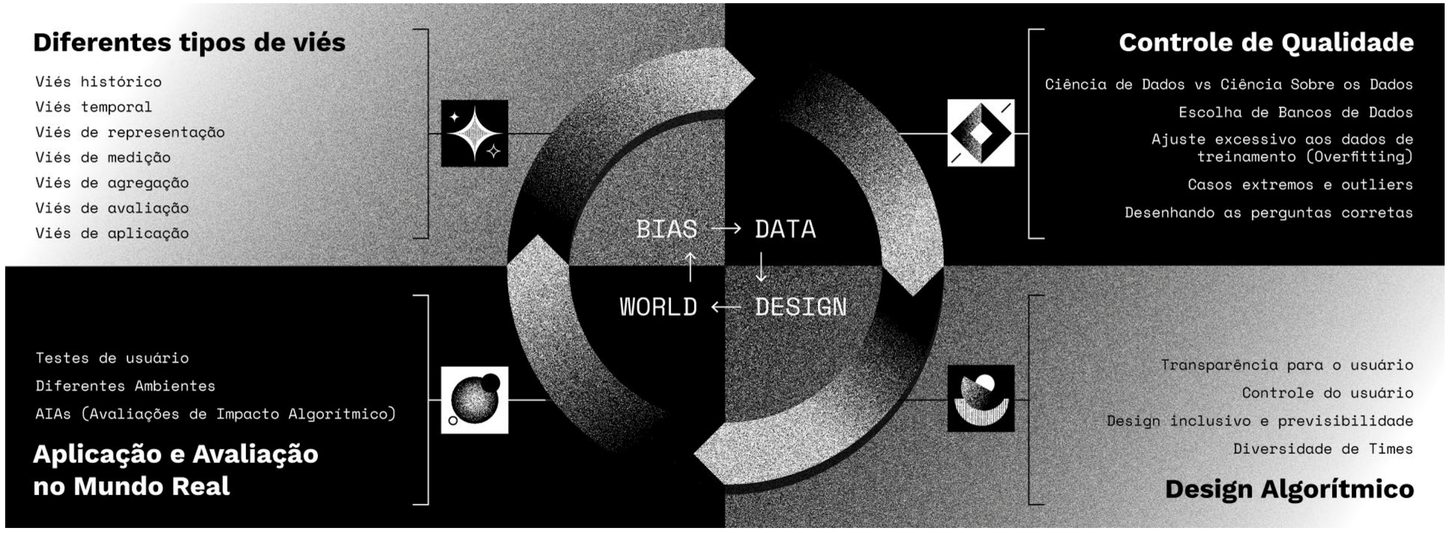
Considere o seguinte cenário: uma pesquisadora médica quer construir um modelo preditivo para ajudar a detectar a presença de câncer de pele. Ela treina o modelo nos registros médicos de um subconjunto de pacientes de um hospital indicando se e quando elas foram diagnosticadas com a doença. Ela observa que o sistema tem uma taxa mais alta de falsos negativos para mulheres negras (existem mais casos de câncer de pele em mulheres negras do que propriamente registrados), e teorizou que o modelo não foi capaz de aprender efetivamente os sinais de câncer de pele nesse conjunto de pacientes por causa da falta de exemplos. Ela procura dados adicionais, representando mulheres negras com câncer de pele, para aumentar o conjunto de dados, retreina o modelo e observa que o desempenho para esta categoria melhora.

Enquanto isso, um colega de trabalho que contrata novos técnicos de laboratório tenta construir um algoritmo para prever a adequação de um candidato a partir de seu currículo, juntamente com escores atribuídos por outras pessoas. Ele percebe que é muito menor a probabilidade de mulheres serem indicadas como candidatas adequadas, do que homens. Assim como sua colega, ele tenta coletar mais amostras de mulheres para adicionar ao conjunto de dados, mas fica desapontado ao ver que o comportamento do modelo não muda.

Por que isso aconteceu? **As fontes dos problemas de desempenho eram diferentes.** No primeiro caso, ela surgiu devido à falta de dados sobre as mulheres negras e introduzir mais dados foi útil. No segundo caso, usar uma avaliação humana de qualidade como um rótulo para estimar os resultados permitiu que o modelo discriminasse por gênero, e coletar mais dados da mesma fonte de distribuição não ajudou.

Muitas tendências humanas podem ser transferidas para as máquinas **porque as tecnologias não são neutras**; elas são tão boas ou ruins, quanto as pessoas que as desenvolvem. Muitas vezes, os algoritmos de aprendizagem de máquina herdam padrões sociais refletidos em seus dados de treinamento, sem qualquer intenção original de programadores para incluir tais tendências. Os cientistas da computação chamam isto de **viés algorítmico**. Viés e imparcialidade são conceitos desafiadores quando se trata de aprendizagem de máquinas, pois não existem métricas claras para o problema. Há uma discordância generalizada até mesmo sobre o que são resultados justos, o que nos gera desafios ainda maiores. Enquanto esses problemas estiverem inseridos na sociedade, nenhum algoritmo estará imparcial.

Para o bem e para o mal, nosso mundo foi transformado pelo Big Data. Para entender os traços digitais gerados por indivíduos, precisamos projetar abordagens multidisciplinares que combinem ciências sociais e ciência de dados. Cientistas sociais e de dados enfrentam o desafio de construir efetivamente sobre as abordagens uns dos outros, para superar as fraquezas de cada lado. Nesta seção, iremos nos aprofundar em algumas fontes potenciais de viés, onde e como os problemas no pipeline de ML podem emergir, e propor um framework para avaliação e mitigação de fontes de riscos nesses pipelines. Identificar e conhecer cada etapa torna as situações menos confusas e mais fáceis de lidar.



Diferentes tipos de viés

Viés histórico

Surge quando há um desalinhamento entre o mundo como ele é, e os valores ou objetivos a serem codificados e propagados em um modelo. Aparece mesmo se os dados forem perfeitamente medidos e amostrados. Um algoritmo, mesmo que reflita o mundo com precisão, ainda pode causar danos à população. As considerações de viés histórico frequentemente envolvem a avaliação do dano representacional (como o reforço de um estereótipo) para um determinado grupo.

Ex.: Algoritmos de NLP que não identificam propriamente o contexto de linguagens de grupos étnicos minoritários (comunidades negras com influência cultural africana são um exemplo).

Viés temporal

Se baseia nos nossos limites de percepção do tempo. Podemos construir um modelo de aprendizado de máquina que funciona bem em um dado momento, mas falha no futuro, porque não levamos em consideração possíveis mudanças futuras ao construir o modelo.

Viés de representação

Dá-se durante a definição e amostragem de uma população. Ocorre quando a população sub representa algum grupo e, subsequentemente, causa pior desempenho para alguma parte da população final. Gera discriminação ou preconceito contra uma pessoa ou grupo por falta de consciência das pessoas desenvolvendo o banco de dados ou o algoritmo. É perigoso porque omite grupos - seja por gênero, raça, deficiência, sexualidade ou classe - causando um apagamento desses. Se um algoritmo com esse tipo de viés é empregado em larga escala, pode reforçar relações de desigualdade social.

Viés de medição

Surge ao escolher, coletar ou computar as métricas e rótulos específicos de interesse. Os recursos considerados relevantes para o resultado são escolhidos, mas podem estar incompletos, serem limitados ou conter ruído, dependendo do grupo ou método de registro. Em muitos casos, a escolha de um único rótulo para criar uma tarefa de classificação pode ser uma simplificação excessiva que direciona o resultado de análises para um valor que não representa a realidade. Também pode surgir do uso de métricas de desempenho que não são granulares ou abrangentes o suficiente.

Ex.: “Credibilidade” é uma construção abstrata, geralmente operacionalizada como uma métrica mensurável, como um escore de crédito. Torna-se problemático, quando os dados são reflexos pobres da construção alvo e/ou são gerados de forma diferente entre os diferentes grupos.

Viés de agregação

Aparece quando suposições errôneas sobre a população afetam a definição do modelo e quando um modelo generalizado é usado para dados nos quais existem valores que devem ser ponderados de forma diferente. Em muitas aplicações, a população de interesse é heterogênea e é improvável que um único modelo atenda a todos os subgrupos. O viés de agregação é uma suposição de que o mapeamento de rótulos é consistente em todos os subconjuntos de dados, mas nem sempre é esse o caso. Um determinado conjunto de dados pode representar pessoas ou grupos com diferentes origens, culturas ou normas, e uma determinada variável pode significar algo bastante diferente entre eles. O viés de agregação pode levar a um modelo que não é ideal para nenhum grupo, ou um modelo que é adequado para uma população restrita (por exemplo, se também houver viés de representação).

Ex.: Análises de Social Media: Algoritmos de processamento natural de linguagem que, ao ser treinado em dados do twitter, não detecta hashtags nem emojis. Ignorar esse contexto específico de grupo em favor de um modelo único e mais geral, construído para todos os dados de mídias sociais, provavelmente levaria a erros de classificação dos tweets dessa população.

Viés de avaliação

Ocorre durante as iterações e avaliações do modelo, quando as populações de teste ou benchmark (comparativas) externas não representam igualmente as várias partes da população de uso. Um modelo é otimizado com seus dados de treinamento, mas sua qualidade é frequentemente medida em benchmarks. (Ex. ImageNet). Esse problema opera em uma escala mais ampla do que outras fontes de viés: um benchmark deturpado incentiva o desenvolvimento e a implantação de modelos que funcionam bem apenas no subconjunto dos dados representados pelos dados desse benchmark. Ocorre devido ao desejo de comparar quantitativamente os modelos uns com os outros. A aplicação de diferentes modelos a um conjunto de bancos de dados externos tenta cumprir este propósito, mas geralmente é estendido para fazer declarações gerais sobre a qualidade de um modelo. Essas generalizações muitas vezes não são estatisticamente válidas e podem levar a um ajuste excessivo a um determinado benchmark (overfitting). Isso é especialmente problemático se o benchmark sofre de viés histórico, de representação ou de medição.

Ex.: Reconhecimento facial comercializado, cujos conjuntos de dados majoritariamente consistem em rostos de homens brancos, causando modelos não apropriadamente treinados para diferentes peles e etnias.

Viés de aplicação/implementação

Surge quando há uma incompatibilidade entre o problema que um modelo se propõe a resolver, e a maneira como ele é realmente usado. Isso geralmente ocorre quando um sistema é construído e avaliado como se fosse totalmente autônomo, quando, na verdade, ele opera em um sistema sociotécnico complicado, moderado por instituições e tomadores de decisão humanos (também chamado de “efeito de enquadramento”). Em alguns casos, os sistemas produzem resultados que devem primeiro ser interpretados por tomadores de decisão humanos. Apesar do bom desempenho isoladamente, eles podem acabar causando consequências prejudiciais por causa de fenômenos como automação e ampliação da escala.

Ex.: Considere um algoritmo usado para informar a decisão de um juiz sobre uma sentença criminal. Um algoritmo que é projetado e

treinado com dados de treinamento de uma jurisdição pode operar incorretamente para outra. Ele pode influenciar as decisões do juiz de maneiras inesperadas e pouco contabilizadas, pois um juiz pode colocar confiança excessiva ou insuficiente no algoritmo, ou até mesmo decidir manter valores contrários aos que se refletem no algoritmo. As consequências para os resultados da justiça criminal, quando um sistema desse tipo é utilizado em contextos complexos, não é clara e pode se tornar algo inesperado ou problemático se uma IA for aplicada sem considerar cenários alternativos.

Controle de qualidade em dados

Ciência de Dados vs Ciência Sobre os Dados

Compreender as várias causas dos vieses é o primeiro passo para a adoção de uma abordagem eficaz. Mas, como os desenvolvedores de algoritmos podem avaliar se seus resultados são, de fato, tendenciosos? Mesmo quando as falhas nos dados de treinamento são corrigidas, os resultados ainda podem ser problemáticos porque o contexto é importante durante a fase de detecção de viés.

Interpretar o que os dados estão realmente falando é diferente do conhecimento estatístico sobre os dados. Dados sozinhos não compõem informação, pois são dependentes do contexto em que estão inseridos. Estar ciente sobre os dados, seus contextos, suas qualidades e fraquezas é essencial ao longo do processo de construção algorítmica. Um algoritmo com o intuito de solucionar problemas socialmente relevantes, se desenvolvido por alguém pouco inserido e distante dos dados que utiliza, dificilmente passará da superfície.

Escolha de Bancos de Dados

A coleta de dados para aprendizado de máquina é um processo delicado e caro. Para a maioria dos algoritmos de ML, os dados de interesse não são apenas “dados pessoais”, mas são quaisquer dados agregados que possam ser relevantes para as operações que serão executadas. Entretanto, o cérebro humano não processa a escala do Big Data, e falha na tentativa de avaliar a integridade de um conjunto de dados muito complexo. O uso de dados disponibilizados publicamente muitas vezes se mostra inconsistente através do conjunto todo e, dependendo da sensibilidade, seu uso pode ser ilegal. Você deve investigar a origem e as características do seu conjunto. Conjuntos de dados de alta qualidade devem ser suficientemente relevantes, representativos, livres de erros e completos, tendo em vista a finalidade pretendida do sistema. Eles também devem ter as propriedades estatísticas adequadas, com equilíbrio/normalização nos dados, incluindo no que diz respeito às pessoas ou grupos de pessoas nas quais o sistema de IA será usado.

Ajuste excessivo aos dados de treinamento (Overfitting)

Acontece quando o modelo de IA pode prever com precisão os valores do conjunto de dados de treinamento, mas não pode prever novos dados com precisão. O modelo adere demais ao conjunto de dados de treinamento, e não se comporta bem para uma população maior e/ou mais diversa. O evento reverso é chamado de underfitting.

Casos extremos e outliers

São dados fora dos limites do conjunto de dados de treinamento, pontos fora da distribuição normal dos dados. Erros e ruídos são classificados como casos extremos: Erros ocorrem quando informações estão ausentes ou são valores incorretos no conjunto de dados; ruído são dados que impactam negativamente o processo de aprendizado de máquina.

Desenhando as perguntas corretas

Um algoritmo é desenhado para responder perguntas, seja um pedido de rótulo para um indivíduo anônimo ou a predição de um cenário futuro. Para garantir a coesão entre a realidade e as simulações, as perguntas pensadas para modelar o algoritmo devem ser claras. Ex.: Em um tipo de modelo de IA chamado “árvore de decisão”, o modelo se parece com uma árvore gigante de perguntas de sim ou não. O objetivo de um dado algoritmo pode ser diferenciar e rotular fotos de cachorros e gatos, mas se o modelo estiver usando perguntas muito específicas, como “a metade da direita da imagem possui mais tons de laranja?”, ele pode até funcionar para um conjunto limitado de dados, mas dificilmente terá sucesso com novas entradas.

Design de algoritmos

Transparência para o usuário

Usuários devem ser capazes de ver e compreender facilmente como seus dados estão sendo coletados e rastreados. Hoje as políticas de cookies, principais dados de rastreio de navegação, normalmente são disponibilizadas em todas as páginas de plataformas, mas existem outras fontes de dados que podem identificar e comprometer um usuário de ponta e elas devem ser esclarecidas.

Controle do usuário

As pessoas devem ter total poder de decisão e ter o direito de escolher se aceitar os serviços de IA, sair de uma interação com um sistema de IA ou interromper sua operação a qualquer momento. O usuário dita o quão personalizado será o serviço, a medida que filtra quais dados ele permitirá coletar. Estas diretrizes atendem às normativas previstas pela LGDP e GDPR.

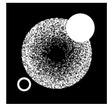
Design inclusivo e previsibilidade

Desenvolvedores e operadores de algoritmos também devem considerar o papel da diversidade em suas equipes de trabalho, nos dados e no nível de sensibilidade cultural em seus processos de tomada de decisão. O emprego da diversidade desde o início do projeto algorítmico

possivelmente evitará efeitos discriminatórios prejudiciais sobre certos grupos, especialmente minorias raciais e étnicas. Embora as consequências imediatas dos vieses nessas áreas possam ser pequenas, a grande quantidade de interações e inferências digitais pode fazer surgir uma nova forma de viés sistêmico. A previsibilidade consiste em prever o impacto que o sistema de IA terá agora e ao longo do tempo. É necessário se perguntar: O uso contínuo e a expansão de escala no meu sistema algorítmico funcionará positivamente, ou pode reforçar padrões de desigualdade entre populações?

Diversidade de Times

O processo de criação de um modelo deve incluir visões e participações de um time diversificado, que construa coletivamente e preencha as diferentes lacunas de conhecimento. Parte da dificuldade de construções diversificadas está relacionada às questões que cada campo considera interessantes, em como percebemos as diferenças e semelhanças das perguntas que fazemos para entender os comportamentos digitais e como nossos métodos podem se complementar. Encontrar questões concretas que sejam relevantes e aplicáveis em diferentes campos é o cerne do processo de design inclusivo. Escutar especialistas que compreendem os contextos éticos de IAs e que representam as vozes dos grupos minoritários também é essencial.



Deployment/ mundo real

Testes de usuário

O teste é uma parte importante da construção de um novo produto ou serviço. O teste do usuário, neste caso, refere-se à obtenção de representantes de diversos grupos que usarão o produto de IA para testá-lo antes de ser lançado e avaliar sua performance.

Diferentes Ambientes

E se o ambiente em que você treinou os dados não for adequado para uma população mais ampla? Exponha seu modelo a ambientes e contextos variados para novos insights. Você precisa ter certeza de que seu modelo pode generalizar para um conjunto mais amplo de cenários.

AIAs (Avaliações de Impacto Algorítmico)

Organizações que empregam sistemas de IA devem rotineiramente coletar feedback sobre as operações correntes, prescrever metodologias para avaliar o impacto que dada operação está tendo sobre os indivíduos envolvidos, quantificar e registrar esses dados, preferencialmente em um report digerível para diferentes setores da empresa e também consumidores/público geral. Esta atividade se refina conforme o nível de maturidade analítica de uma organização se desenvolve, e ela aprende a conhecer seus algoritmos e seus dados tão bem quanto conhece seus funcionários. Essas avaliações são mais efetivas quando efetuadas por um time diverso, composto por pessoas que representam o público alvo final.

É importante que os operadores e desenvolvedores de algoritmos sempre se perguntem: **deixaremos alguns grupos de pessoas em pior situação como resultado do design do algoritmo ou de suas consequências indesejadas?**

O advento da IA nos apresenta a oportunidade de, com mudanças sociotécnicas, poder ajudar a trazer um mundo melhor e mais justo. Criar um algoritmo não é apenas descrever uma fórmula ou código computacional, mas é criar um legado tecnológico com impacto social real. Portanto, uma visão humanística ao longo do processo de design algorítmico é essencial para um legado positivo, permitindo assim a criação de um futuro melhor.

05

papéis e responsabilidades

Governança, políticas e responsabilidade

05

papéis e responsabilidades

Quando ouvimos o termo governança, somos automaticamente transportados para um sentimento de obrigatoriedade e conservadorismo. No entanto, com o volume de dados que geramos diariamente, precisamos buscar uma mudança de mindset com relação ao assunto, que passa a ser uma camada de proteção para nós e de avanços em alguns temas para as empresas.
Governança é um termo bastante amplo porque se relaciona com diferentes assuntos dentro de uma organização, que podem englobar desde questões comportamentais como, por exemplo, o dress code, até questões legais, como o que fazer se descobrir um colega cometendo fraude.

Como compartilhamos muitos dados pessoais e estamos expostos a muitas ferramentas que tomam decisões por nós, consequentemente muitas empresas possuem acesso aos nossos dados, se tornando guardiãs dos mesmos, além de responsáveis pelos impactos que seus produtos geram nas pessoas e sociedade.

Empresas que fazem uso de dados e de Inteligência Artificial devem definir claramente quais questões éticas são inegociáveis e torná-las assunto primordial. Um dos caminhos é incluir essas definições em uma governança do uso de dados e comunicar estas decisões de forma transparente. Muitas organizações estão formando comitês multidisciplinares de governança de algoritmos como forma de mostrar a importância do tema e para evitar que este assunto continue sendo atribuído como responsabilidade das áreas de tecnologia da informação (TI).

A governança no uso de dados e algoritmos deve responder perguntas do tipo: Como a empresa lida com os dados de seus colaboradores, seus parceiros e seus clientes? Quais procedimentos adota para evitar que suas ferramentas e produtos sejam enviesados? Quais definições éticas são inegociáveis e devem fazer parte do desenvolvimento de qualquer ferramenta analítica? Como a área técnica pode comprovar que está seguindo um padrão ético? Esses são exemplos de questões que podem fazer parte da governança e englobam políticas e processos e dependem da estrutura escolhida pela empresa para gerenciar dados.

Desde o escândalo da Cambridge Analytica, houve um aumento da conscientização sobre privacidade e mau uso dos dados, o que fortaleceu a implantação de leis de proteção de dados por parte de órgãos reguladores. As empresas precisaram se adequar a essas novas leis - a General Data Protection Regulation (GDPR), na União Européia, e a Lei Geral de Proteção de Dados (LGPD), no Brasil - que exigem uma série de adequações para garantir a privacidade e segurança das pessoas. São exemplos: a obrigatoriedade de solicitar aos usuários permissão de coleta de dados nos websites, a disponibilização de opções de quais dados o usuário permite coleta e ainda a garantia de deleção de vínculo de dados que permitam identificar o indivíduo quando solicitado e em casos de finalização do relacionamento.

Além disso, seguir processos éticos também pode fazer com que a empresa tenha menos desperdício de dinheiro, ganhe mais confiança de seus clientes, colaboradores e parceiros e seja mais efetiva em processos de contratação e compras. Uma boa governança contribui para a imagem da empresa e saber como comunicar sobre essa questão - tanto internamente quanto externamente - é essencial.

Os pontos principais para uma boa governança são:

- As áreas e setores da organização precisam conhecer e estar alinhadas com a governança de dados estabelecida pela empresa;
- Os colaboradores precisam de treinamentos para entender como o uso dos dados e das técnicas são e podem ser utilizados na empresa e potencializar o seu trabalho;
- A responsabilidade pelas ferramentas ultrapassa o momento da venda do produto. A empresa precisa estar preparada para prestar contas desde a concepção até o produto final.

06

sistemas éticos para IA

Ricardo Cappra // 17 de novembro de 2021

06

sistemas éticos para IA

Sempre que julgamos um sistema tecnológico por seu posicionamento ético, é fundamental revisitar o processo de concepção do mesmo. Uma série de premissas e regras pré-estabelecidas, associadas ao ambiente e cultura em que o mesmo está instalado, complementado por rotinas contínuas de controle, atualização e suporte, fazem com que as normas de conduta formem, originalmente, o comportamento ativo daquele sistema.

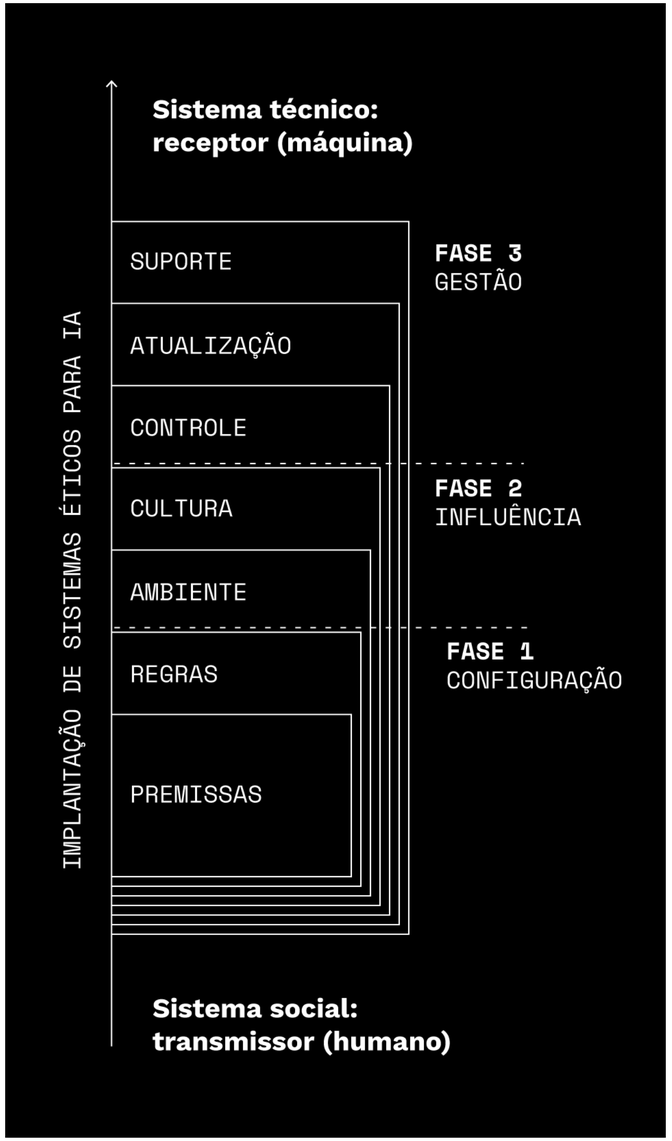
Se observarmos com atenção essa fusão de componentes, logo perceberemos que ocorre o mesmo com nosso sistema social formado por humanos. A partir do momento que determinado indivíduo recebe configurações originais por parte de sua família, escola, bairro, amigos, redes, ele passa a atualizar suas premissas éticas conforme a evolução do seu modo de viver.

A máquina, quando caracterizada como Inteligência Artificial, representa um repositório de informações e normas dos indivíduos que a configuraram, tornando-se, assim, uma forma de sistema autônomo baseado nas informações daqueles que a abasteceram. A autonomia, como conceito, é encontrada nos debates sobre moral e costuma ser definida como a capacidade racional de tomar decisões não forçadas baseando-se nas informações disponíveis. A premissa de autonomia não significa que não foram carregados traços históricos na definição das regras de funcionamento, afinal ele é formado por determinado meio, geralmente respeitando os limites éticos dos seus formadores. Talvez a grande diferença esteja na capacidade crítica que o ser humano possui de, a partir de sua própria percepção, alterar um determinado comportamento espontaneamente para que este esteja mais adequado às suas crenças de viver bem. Já a máquina dependerá da alteração de um determinado código para, assim, possuir um novo padrão de ação.

Se transportarmos isso para dentro do mundo corporativo, será necessário observar as pessoas que fazem parte do sistema social no ambiente de trabalho, geralmente formado pelo comportamento dos colaboradores que fazem parte daquele negócio. A partir daí, será possível identificar os principais parâmetros éticos que estão ali presentes. Caso essa análise prévia não seja realizada, corre-se o risco da criação de sistemas com divergências éticas, que serão inseridas no sistema tecnológico a ser desenvolvido, transferindo assim uma inconsistência, ou desvio ético, para uma rotina automatizada. Para exemplificar, em um sistema social com pouca diversidade, a inserção de códigos nos softwares para que promovam um processo seletivo mais diverso pode gerar uma série de problemas. A tentativa de inserir nas máquinas um padrão de comportamento não adotado por aqueles indivíduos, está carregada de elementos divergentes, algo que inevitavelmente irá incorrer em uma falha sistêmica, seja ela social ou técnica. A falha, nesse caso, é no sistema social nativo, e não no código inserido para ativação por meio de Inteligência Artificial.

Em razão disso, um comitê para compreender os princípios de comportamento e estabelecer premissas, que acompanhará a implementação das mesmas em forma de códigos, é algo fundamental em uma era onde máquinas interagem continuamente com humanos. Ele irá testar as hipóteses e possíveis falhas, além de monitorar a contínua evolução dessa tecnologia enquanto ativa no sistema social. Aqui a premissa de um modelo de governança estruturado, que dialogue desde a fase de concepção, até a execução desse sistema inteligente, quando o mesmo torna-se um agente ativo da própria organização, é fundamental. Sistemas tecnológicos são reprodutores de tarefas, sejam aquelas definidas como premissa do aprendizado de máquina ou os resultados de tarefas executadas pelos mesmos.

Sistemas éticos, sejam eles tecnológicos ou sociais, são formados por componentes similares: premissas, regras, ambiente, cultura, controle, atualização e suporte, sendo que quando essas partes não são devidamente supervisionadas e integradas, eleva-se o risco de uma falha comportamental desses autônomos. Então, na construção de um sistema de Inteligência Artificial é fundamental uma atenção redobrada com relação aos aspectos éticos e na influência que os ambientes sociais podem gerar enquanto esses sistemas estiverem em funcionamento. Lembrando que a responsabilidade do comportamento dessas máquinas é inteira dos seus criadores.



07

mergulhe no assunto

Analisando a ética de aplicações reais

A história antiética da ciência de dados na política

A Simulmatics é talvez a primeira empresa do ramo computacional que atuou na política. Com a ideia de que, se pudessem coletar dados suficientes sobre pessoas suficientes e escrever códigos bons o suficiente, qualquer cenário poderia ser simulado e previsto, e então manipulado com mensagens direcionadas, empregou uma equipe de cientistas de dados brilhantes (e com conhecimento em guerra psicológica) para atuar na campanha presidencial de John F. Kennedy. As limitações tecnológicas da época impediram seus objetivos, mas agora o caso é estudado como precursor de situações e práticas escandalosas hoje dentro da indústria.

LEIA AQUI

A coleta de dados que moldaram a política moderna

Em março de 2018, o Facebook admitiu a transferência ilegal de milhões de perfis de usuários para a empresa de análise de dados Cambridge Analytica. A empresa contratada pela campanha presidencial de Donald Trump coletou informações privadas de aproximadamente 270.000 usuários, e suas redes de amigos, sob falsos pretextos de um aplicativo voltado para pesquisa. Todos os usuários que participaram da pesquisa consentiram em ter seus dados coletados, mas foram informados que era para “uso acadêmico”. O aplicativo subsequentemente rastreou os dados dos amigos desses usuários sem seu conhecimento ou consentimento, e os transferiu para a Cambridge Analytica. O número de atingidos chegou a até 87 milhões de usuários, tornando-se uma das maiores transferências de dados ilegais na história. Estes dados foram utilizados para criar um perfil “psicossocial” dos usuários da rede, e assim direcionar anúncios personalizados que apelassem para os medos e inseguranças identificados para cada perfil, a fim de influenciar as escolhas políticas e de voto destes.

LEIA AQUI

Resisting the rise of facial recognition

A implementação de câmeras de reconhecimento facial tem se expandido ao redor do mundo, principalmente por causa da popularização das chamadas smart cities. Com essa expansão surgem também vários problemas relacionados a seu uso, como a falta de consentimento das pessoas, já que registrar os rostos e comparar com um banco e armazenar as informações de onde e quando cada pessoa esteve é bem diferente de ter uma câmera na rua gravando o que aconteceu no local, como acontecia até então. Além disso, muitas pesquisas têm mostrado que essas tecnologias falham principalmente para pessoas negras e latinas, já que não houve uma preocupação das fabricantes em treinar os modelos para reconhecer diversos tipos de rostos e peles, o que tem resultado em muitos casos de pessoas apreendidas erroneamente.

LEIA AQUI

Recrutamento automático, e enviesado, de funcionários

Em 2014, a varejista Amazon, cuja força de trabalho global é 60% masculina e onde os homens detêm 74% dos cargos gerenciais da empresa, recentemente interrompeu o uso de um algoritmo de recrutamento após descobrir o preconceito de gênero. Os dados que os engenheiros usaram para criar o algoritmo foram derivados de currículos enviados à empresa ao longo de um período de 10 anos, que eram predominantemente de homens brancos. O algoritmo foi ensinado a reconhecer padrões de palavras nos currículos, ao invés de conjuntos de habilidades relevantes, e esses dados foram comparados com o departamento de engenharia predominantemente masculino da empresa para determinar a adequação de um candidato. Como resultado, o software de IA penalizou qualquer currículo que contivesse a palavra “feminino” no texto e rebaixou os currículos de mulheres, resultando em preconceito de gênero. Os recrutadores da Amazon tiveram de abandonar o software de avaliação devido a essas questões de discriminação e justiça.

LEIA AQUI

Predição de risco criminal

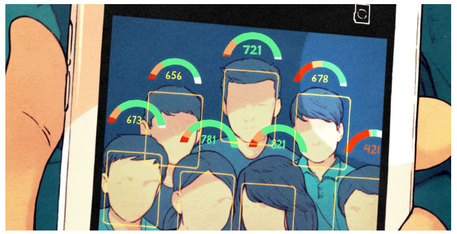
Nos últimos anos, tem aumentado o uso de ferramentas de avaliação de risco no Sistema de Justiça Criminal dos EUA, principalmente para definição se acusados irão aguardar o julgamento em liberdade, mas são utilizados em diversas fases, inclusive para definição de pena e alocação de presos dentro de prisões. Essas ferramentas informam aos juízes se o acusado tem alto risco de reincidir e com base nisso o juiz decide se a pessoa vai aguardar o julgamento em liberdade (cumprir mais tempo de pena, sair em condicional, etc). O problema é que essas ferramentas utilizam dados do sistema de justiça criminal, que já criminaliza desproporcionalmente pessoas pobres e não-brancas, e, em muitos casos, informações financeiras, geográficas, empregatícias e até comportamentais para avaliar o risco.

LEIA AQUI

As limitações tecnológicas para peles mais escuras

A pesquisadora do MIT, Joy Buolamwini, descobriu que algoritmos que alimentam três sistemas de software de reconhecimento facial comercializados não reconheciam peles mais escuras. Ainda, no geral, a maioria dos conjuntos de dados de treinamento de reconhecimento facial são estimados em mais de 75% de sexo masculino e mais de 80% de pele branca em suas composições. De acordo com a pesquisa de Buolamwini, as taxas de erro para os três produtos eram menos de um por cento no geral, mas aumentaram para mais de 20% em um produto e 34% nos outros dois na identificação de mulheres de pele mais escura. Em resposta às descobertas, tanto a IBM quanto a Microsoft se comprometeram a melhorar a precisão de seu software de reconhecimento para rostos de pele mais escura.

LEIA AQUI



Um panóptico social

Desde 2014, o governo chinês vem implantando um sistema de crédito social. Os cidadãos recebem um escore e vão ganhando ou perdendo pontos dependendo das coisas que fazem e são penalizados se tem poucos pontos como, por exemplo, não conseguir comprar uma passagem de avião. As informações saem de outros sistemas que estão interligados, como sistema de compras ou de reconhecimento facial. Além disso, esse sistema faz parte de um grande sistema de vigilância que a China está implementando, que tem vários outros problemas também, como a falta de transparência - não se sabe quais dados são coletados, como são usados, qual cálculo, etc - ou utilização de dados incorretos e questões relacionadas à privacidade e abuso de poder.

LEIA AQUI

Governança ética dentro de Big Techs

Timnit Gebru, pesquisadora de ramificações sociais e éticas da Inteligência Artificial do Google, conta que seu gerente pediu que ela retirasse o nome de um artigo de pesquisa de sua co-autoria, porque uma revisão interna considerou o conteúdo questionável. Ela acabou sendo demitida sem aviso prévio por conta disto.
O artigo discute questões éticas levantadas por avanços recentes na tecnologia de IA que funciona com a linguagem que o Google disse ser importante para o futuro de seus negócios.
O artigo examina pesquisas anteriores sobre as limitações dos sistemas de IA que analisam e geram linguagem, e cita estudos que mostram que esses algoritmos de NLP podem consumir grandes quantidades de eletricidade e amplificar preconceitos encontrados em textos online. Eles sugerem maneiras pelas quais os pesquisadores de IA podem ser mais cuidadosos com a tecnologia, inclusive documentando melhor os dados usados para criar tais sistemas. Depois do ocorrido, 2.200 funcionários do Google assinaram uma carta demandando mais transparência e o caso levantou muitas questões sobre os papéis de governança das Big Techs.

LEIA AQUI

Um problema aparentemente inofensivo

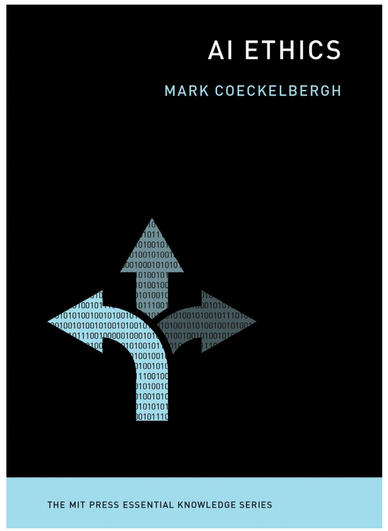
As famosas pulseiras de monitoramento cardíaco, de empresas como FitBIt e Samsung, têm dificuldade de captar o monitoramento cardíaco de pessoas com a pele mais escura, pois as fabricantes optam por utilizar a luz verde, mais barata, para identificar os batimentos cardíacos. Ao contrário da luz infravermelha, a luz verde tem dificuldade de reconhecer o volume de sangue em peles com mais melanina, ou seja, peles mais escuras. Esse caso mostra que as escolhas feitas pelas empresas podem excluir grupos de pessoas, mesmo que essa exclusão pareça inofensiva. Também é necessário discutir que nenhuma dessas empresas se importaram em fazer testes em outros tipos de peles antes de lançar o produto.

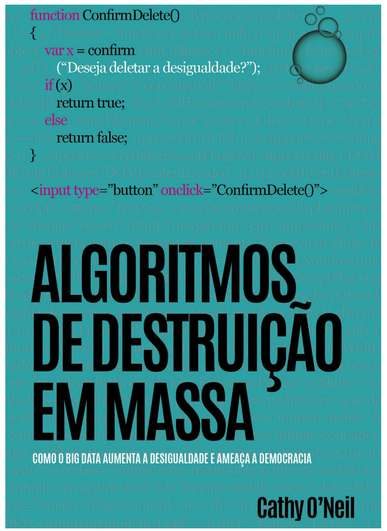
LEIA AQUI

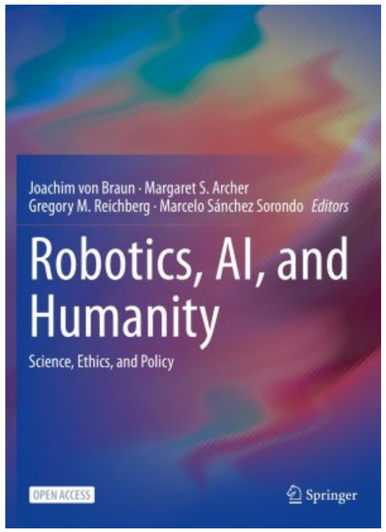
Carros autônomos e a discussão de responsabilização

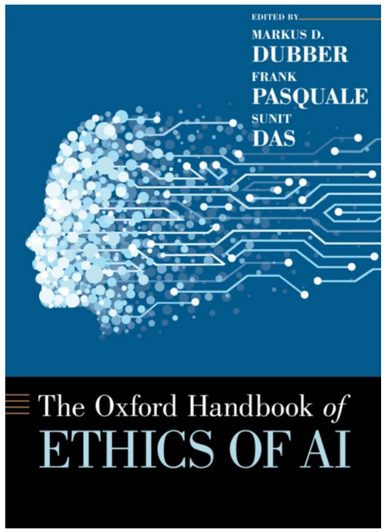
Outra tecnologia que tem gerado polêmica é a dos carros autônomos. Dois acidentes fatais recentes, um envolvendo a empresa Uber, em 2020, e outro a Tesla, em 2021, fizeram reacender a discussão. A primeira morte por um atropelamento envolvendo um carro autônomo foi em 2018 e desde então há várias disputas na justiça, principalmente em torno de quem é o responsável pelos acidentes. Isso ocorre pois essa tecnologia ainda está em teste e por esse motivo utilizam motoristas humanos em algumas ações de controle do veículo. Dessa forma, as empresas procuram sempre jogar a culpa para o motorista humano para poupar os erros da tecnologia, mesmo quando são eles o motivo do acidente.

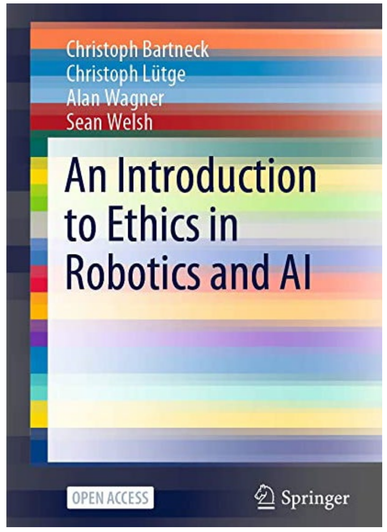
LEIA AQUI **E AQUI**

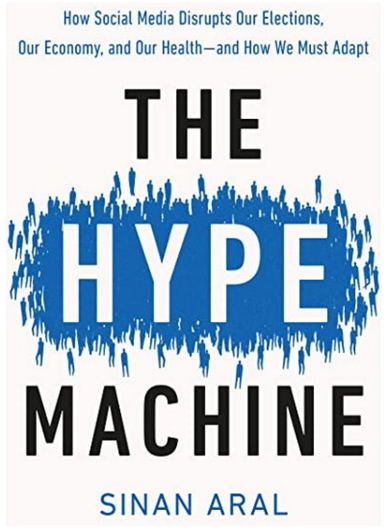

AI ETHICS
MARK COECKELBERGH
THE MIT PRESS ESSENTIAL KNOWLEDGE SERIES

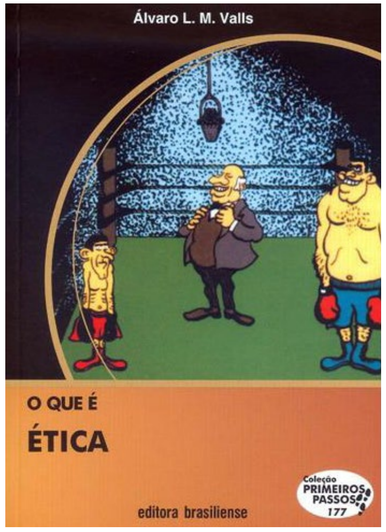

function ConfirmDelete()
{
var x = confirm
("Deseja deletar a desigualdade?");
if (x)
return true;
else
return false;
}
<input type="button" onclick="ConfirmDelete()">
ALGORITMOS DE DESTRUIÇÃO EM MASSA
COMO O BIG DATA AUMENTA A DESIGUALDADE E AMEAÇA A DEMOCRACIA
Cathy O'Neil

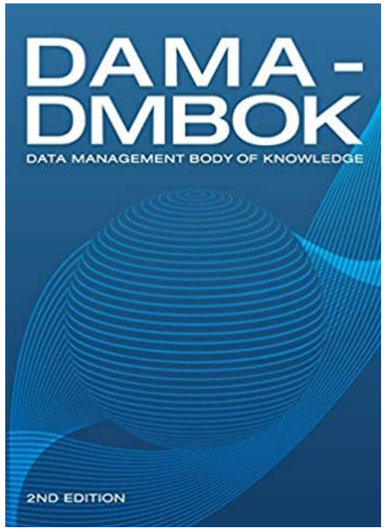

Joachim von Braun · Margaret S. Archer
Gregory M. Reichberg · Marcelo Sánchez Sorondo Editors
Robotics, AI, and Humanity
Science, Ethics, and Policy
OPEN ACCESS
Springer


EDITED BY
MARKUS D. DUBBER
FRANK PASQUALE
SUNIT DAS
The Oxford Handbook of
ETHICS OF AI


Christoph Bartneck
Christoph Lütge
Alan Wagner
Sean Welsh
An Introduction to Ethics in Robotics and AI
OPEN ACCESS
Springer


How Social Media Disrupts Our Elections,
Our Economy, and Our Health—and How We Must Adapt
THE HYPE MACHINE
SINAN ARAL


Álvaro L. M. Valls
O QUE É
ÉTICA
editora brasiliense
Coleção PRIMEIROS PASSOS 177


DAMA-DMBOK
DATA MANAGEMENT BODY OF KNOWLEDGE
2ND EDITION

referências

Ada Lovelace Institute, AI Now Institute and Open Government Partnership. (2021). Algorithmic Accountability for the Public Sector. Available at: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

Adam, Alison. "Computer Ethics in a Different Voice." Information and Organization, vol. 11, no. 4, Oct. 2001, pp. 235–61. DOI.org (Crossref), https://doi.org/10.1016/S1471-7727(01)00006-9.

Ananny, Mike. "Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness." Science, Technology, & Human Values, vol. 41, no. 1, Jan. 2016, pp. 93–117. SAGE Journals, https://doi.org/10.1177/0162243915606523.

Asadi Someh, Ida, et al. "ETHICAL IMPLICATIONS OF BIG DATA ANALYTICS." Research-in-Progress Papers, June 2016, https://aisel.aisnet.org/ecis2016_rip/24.

Bommasani, Rishi, et al. "On the Opportunities and Risks of Foundation Models." ArXiv:2108.07258 [Cs], Aug. 2021. arXiv.org, http://arxiv.org/abs/2108.07258.

Broad, Terence, et al. "Active Divergence with Generative Deep Learning - A Survey and Taxonomy." ArXiv:2107.05599 [Cs], July 2021. arXiv.org, http://arxiv.org/abs/2107.05599

Brown, Shea, et al. "The Algorithm Audit: Scoring the Algorithms That Score Us." Big Data & Society, vol. 8, no. 1, Jan. 2021, p. 2053951720983865. SAGE Journals, https://doi.org/10.1177/2053951720983865.

Christiano, Paul, et al. "Deep Reinforcement Learning from Human Preferences." ArXiv:1706.03741 [Cs, Stat], July 2017. arXiv.org, http://arxiv.org/abs/1706.03741.

De-Arteaga, Maria, et al. "Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting." Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 2019, pp. 120–28. ACM Digital Library, https://doi.org/10.1145/3287560.3287572.

Drew, Cat. "Data Science Ethics in Government." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2083, Dec. 2016, p. 20160119. royalsocietypublishing.org (Atypon), https://doi.org/10.1098/rsta.2016.0119.

Floridi, Luciano, and Mariarosaria Taddeo. "What Is Data Ethics?" Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2083, Dec. 2016, p. 20160360. royalsocietypublishing.org (Atypon), https://doi.org/10.1098/rsta.2016.0360.

Firth-Butterfield, Kay. "Building an Organizational Approach to Responsible AI", MITSloan, https://sloanreview.mit.edu/article/building-an-organizational-approach-to-responsible-ai/

Gadekallu, Thippa Reddy, et al. "Federated Learning for Big Data: A Survey on Opportunities, Applications, and Future Directions." ArXiv:2110.04160 [Cs], Oct. 2021. arXiv.org, http://arxiv.org/abs/2110.04160.

Ganapini, Marianna. "Exploring the Under-Explored Areas in Teaching Tech Ethics Today." Montreal AI Ethics Institute, 21 July 2021, https://montrealethics.ai/exploring-the-under-explored-areas-in-teaching-tech-ethics-today/.

Gender Shades. http://gendershades.org/. Accessed 3 Nov. 2021.

Halevy, Matan, et al. "Mitigating Racial Biases in Toxic Language Detection with an Equity-Based Ensemble Framework." ArXiv:2109.13137 [Cs], Sept. 2021. arXiv.org, https://doi.org/10.1145/3465416.3483299.

Henighan, Tom, et al. "Scaling Laws for Autoregressive Generative Modeling." ArXiv:2010.14701 [Cs], Nov. 2020. arXiv.org, http://arxiv.org/abs/2010.14701.

Metcalf, Jacob, Emanuel Moss, Elizabeth Anne Watkins, et al. Algorithmic Impact Assessments and Accountability: The Co-Construction of Impacts. SSRN Scholarly Paper, ID 3736261, Social Science Research Network, 29 Sept. 2020. papers.ssrn.com, https://papers.ssrn.com/abstract=3736261.

Metcalf, Jacob, Emanuel Moss, and danah boyd. "Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics." Social Research: An International Quarterly, vol. 86, no. 2, 2019, pp. 449–76.

Practical Data Ethics | Data Ethics. https://ethics.fast.ai/. Accessed 3 Nov. 2021.

"Privacy in the Brain: The Ethics of Neurotechnology." Neuroscience from Technology Networks, https://www.technologynetworks.com/neuroscience/articles/privacy-in-the-brain-the-ethics-of-neurotechnology-353075. Accessed 3 Nov. 2021.

Ryan, Mark. "In AI We Trust: Ethics, Artificial Intelligence, and Reliability." Science and Engineering Ethics, vol. 26, no. 5, Oct. 2020, pp. 2749–67. Springer Link, https://doi.org/10.1007/s11948-020-00228-y.

---. **"In AI We Trust: Ethics, Artificial Intelligence, and Reliability."** Science and Engineering Ethics, vol. 26, no. 5, Oct. 2020, pp. 2749–67. Springer Link, https://doi.org/10.1007/s11948-020-00228-y.

Saltz, Jeffrey, et al. "Integrating Ethics within Machine Learning Courses." ACM Transactions on Computing Education, vol. 19, no. 4, Aug. 2019, p. 32:1-32:26. November 2019, https://doi.org/10.1145/3341164.

Saltz, Jeffrey S., et al. "Key Concepts for a Data Science Ethics Curriculum." Proceedings of the 49th ACM Technical Symposium on Computer Science Education, Association for Computing Machinery, 2018, pp. 952–57. ACM Digital Library, https://doi.org/10.1145/3159450.3159483.

Schiff, Daniel, et al. "Explaining the Principles to Practices Gap in AI." IEEE Technology and Society Magazine, vol. 40, no. 2, June 2021, pp. 81–94. IEEE Xplore, https://doi.org/10.1109/MTS.2021.3056286.

Stahl, Bernd Carsten, et al. "The Ethics of Computing: A Survey of the Computing-Oriented Literature." ACM Computing Surveys, vol. 48, no. 4, Feb. 2016, p. 55:1-55:38. May 2016, https://doi.org/10.1145/2871196.

Suresh, Harini, and John V. Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle." ArXiv:1901.10002 [Cs, Stat], June 2021. arXiv.org, http://arxiv.org/abs/1901.10002

Taddeo, Mariarosaria. "Data Philanthropy and the Design of the Infraethics for Information Societies." Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2083, Dec. 2016, p. 20160113. royalsocietypublishing.org (Atypon), https://doi.org/10.1098/rsta.2016.0113.

Taylor, Linnet. “The Ethics of Big Data as a Public Good: Which Public? Whose Good?” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2083, Dec. 2016, p. 20160126. royalsocietypublishing.org (Atypon), https://doi.org/10.1098/rsta.2016.0126.

“The Chief AI Ethics Officer: A Champion or a PR Stunt?” Montreal AI Ethics Institute, 17 May 2021, https://montrealethics.ai/the-chief-ai-ethics-officer-a-champion-or-a-pr-stunt/.

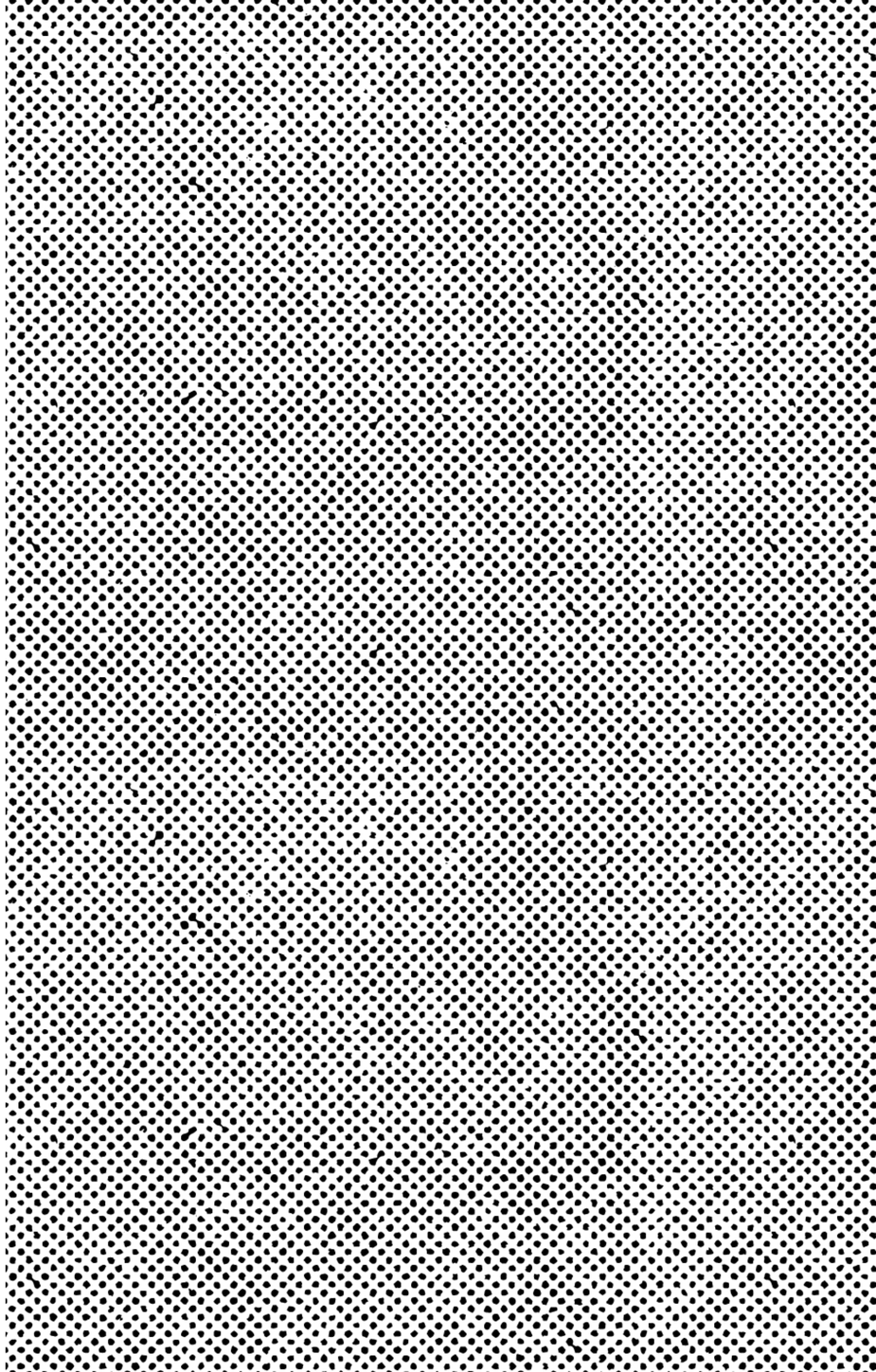
Thomas, Rachel, and David Uminsky. “The Problem with Metrics Is a Fundamental Problem for AI.” ArXiv:2002.08512 [Cs], Feb. 2020. arXiv.org, http://arxiv.org/abs/2002.08512.

Varley-Winter, Olivia, and Hetan Shah. “The Opportunities and Ethics of Big Data: Practical Priorities for a National Council of Data Ethics.” Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2083, Dec. 2016, p. 20160116. royalsocietypublishing.org (Atypon), https://doi.org/10.1098/rsta.2016.0116.

Vlasceanu, M., Dudik, M., & Momennejad, I. (2021, August 29). Network Structure, Gender Diversity, and Interdisciplinarity Predict the Centrality of AI Organizations. https://doi.org/10.31234/osf.io/dp3ef

Warning Signs: The Future of Privacy and Security in an Age of Machine Learning. https://iapp.org/resources/article/warning-signs-the-future-of-privacy-and-security-in-an-age-of-machine-learning/. Accessed 3 Nov. 2021.

Whittlestone, Jess, and Jack Clark. “Why and How Governments Should Monitor AI Development.” ArXiv:2108.12427 [Cs], Aug. 2021. arXiv.org, http://arxiv.org/abs/2108.12427.



Cappra Institute for Data Science

O Cappra Institute é um centro de pesquisa independente e privado, com sede nos EUA, que reúne especialistas de diferentes partes do mundo para estudar as interseções entre dados, humanos e organizações. Para promover a cultura analítica, o instituto realiza análises transdisciplinares que explicam os impactos da tecnologia da informação no mundo e cria métodos para facilitar a jornada de profissionais e empresas na era Big Data.

Entre as atividades realizadas pelo Cappra Institute estão: produção de pesquisas transdisciplinares, geração de conhecimento acessível, criação de metodologias analíticas, realização de eventos especializados, preparação de formadores e especialistas, realização de iniciativas de cultura analítica e promoção de melhores práticas data-driven. Entre as empresas e organizações que já se beneficiaram diretamente das iniciativas realizadas e dos métodos desenvolvidos pelo Cappra Institute estão: Governo dos Estados Unidos, Banco Mundial, Unilever, Santander, Volvo, Abbott Laboratories, Whirlpool, UOL, Ambev/AB Inbev, entre tantas outras.

CAPPRA
INSTITUTE
FOR DATA
SCIENCE

Cappra Institute (Cappra LLC)
601 Brickell Key Drive, Suite 901
Miami FL 33131 US
www.cappra.institute

Editorial

Ane Schutz

Guilherme Machado

Iara Passos

Iohana Bernardes

Júlia Bergmann

Léo Ambros

Mariana Mizutani

Ricardo Cappra