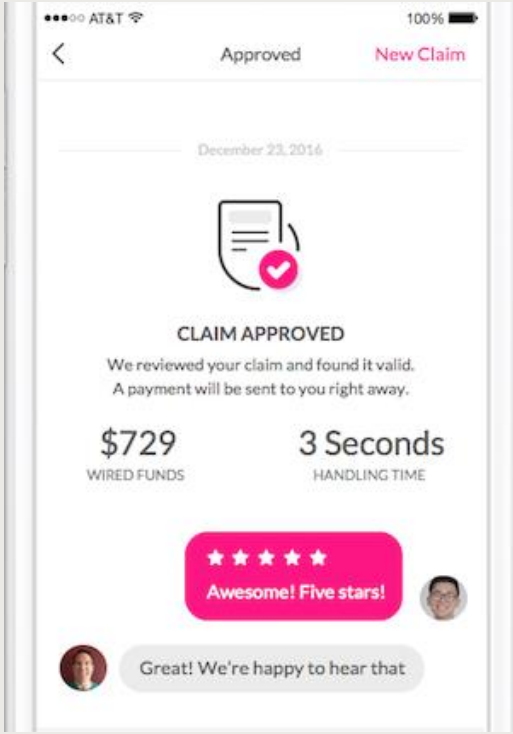
Fairness and Model Explainability

Fairness

“A computer system might be considered biased if it discriminates against certain individuals or groups of individuals.”

https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-detect-data-bias.html

AI in Action - Insurance



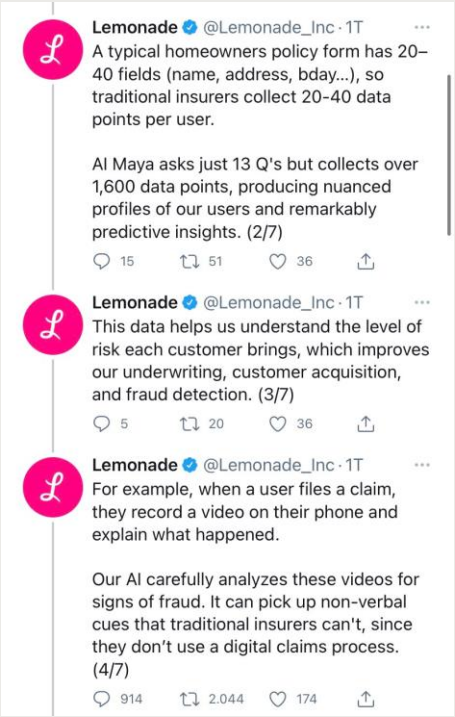
[Fastest insurance claim paid out to a customer: world record set by Lemonade's claims bot AI Jim]

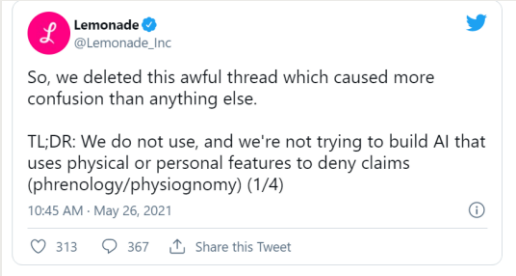
"reviewed a customer's claim, cross referenced it with the policy, ran 18 anti-fraud algorithms on it, approved the claim, sent wiring instructions to the bank, and informed the client the claim was closed - all within three seconds and zero paperwork"

Accusations of Bias and Discrimination

"Lemonade tweeted about what it means to be an AI-first insurance company. It left a sour taste in many customers' mouths"

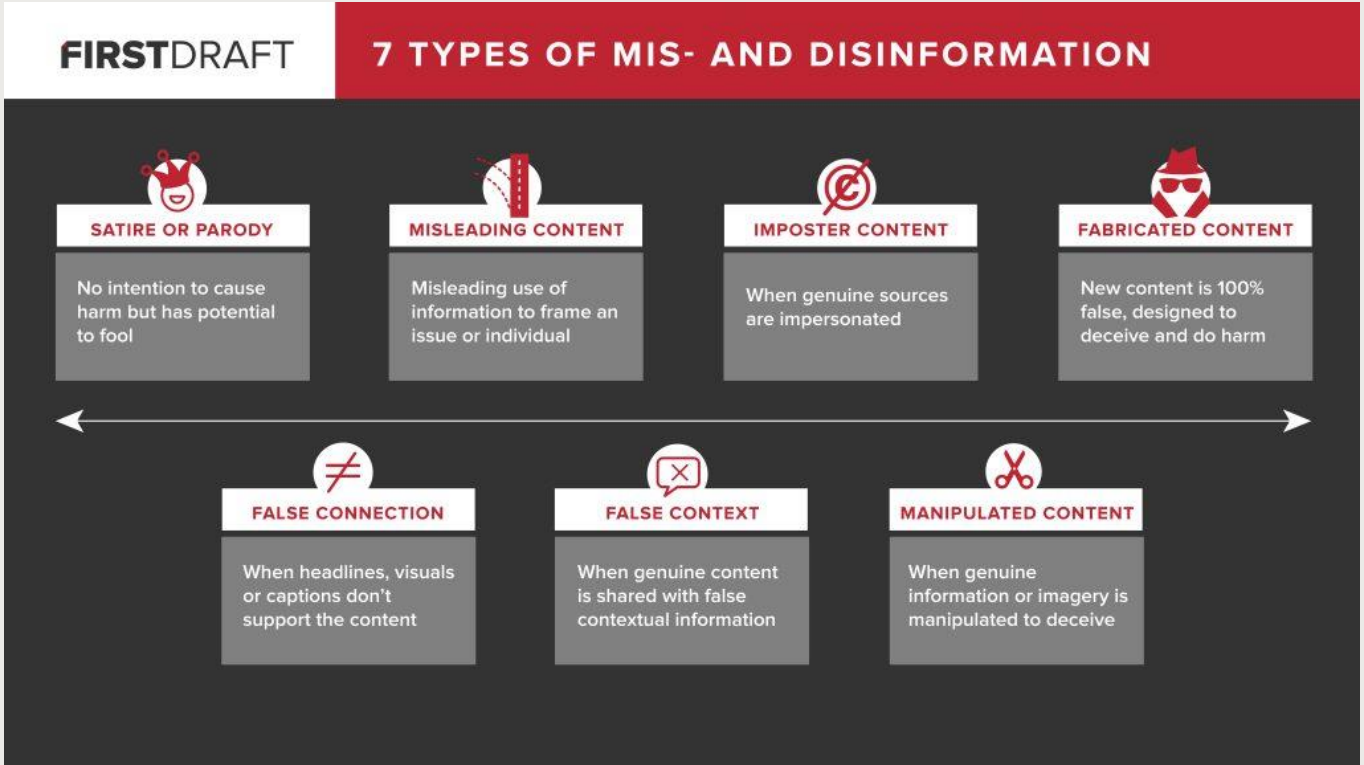
https://www.vox.com/recode/22455140/lemonade-insurance-ai-twitter





Fake News on Social Media

Fake news. It’s complicated. By Claire Wardle
https://firstdraftnews.org/articles/fake-news-complicated/



AI in Action – Social Media

Twitter promises to fine-tune its 5G coronavirus labeling after unrelated tweets were flagged

Tweets with the words "oxygen" and "frequency" were being tagged with a fact-check label

https://www.theverge.com/2020/6/27/21305503/twitter-labels-5g-conspiracy-coronavirus

AI in Action – Image Cropping

Twitter says its image-cropping algorithm was biased, so it's ditching it

“when tested on randomly linked images of people of various races and genders, favored White people over Black people and women over men”

https://www.cnn.com/2021/05/19/tech/twitter-image-cropping-algorithm-bias/index.html

AI in Action – Credit Card

The Apple Card Didn't 'See' Gender—and That's the Problem

- *“users noticed that it seemed to offer smaller lines of credit to women than to men”*
- *“No one from the company seemed able to describe how the algorithm even worked, let alone justify its output.”*

https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

Algorithm is gender-blind

“Goldman landed on what sounded like an ironclad defense: The algorithm, it said, has been vetted for potential bias by a third party; moreover, it doesn’t even use gender as an input. How could the bank discriminate if no one ever tells it which customers are women and which are men?”

https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

Proxies

The idea that removing an input eliminates bias is "*a very common and dangerous misconception*," says Rachel Thomas, a professor at the University of San Francisco and the cofounder of Fast.ai, a project that teaches people about AI

https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

Rachel Thomas – Medium

medium.com/@racheltho?p=909a406e4a71

Jan 30, 2019

I'm an AI researcher, and here's what scares me about AI

Rachel Thomas

fast.ai co-founder & professor USF Data Institute | twitter: @math_rachel

Follow

1.5K 8

AI is being increasingly used to make important decisions. Many AI experts (including Jeff Dean, head of AI at Google, and Andrew Ng, founder of Coursera and deeplearning.ai) say that warnings about sentient robots are overblown, but other harms are not getting enough attention. I agree. I am an AI researcher, and I'm worried about some of the societal impacts that we're already seeing. In particular, these 5 things scare me about AI:

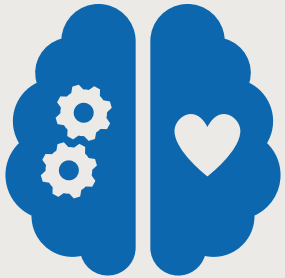
1. Algorithms are often implemented without ways to address mistakes.
2. AI makes it easier to not feel responsible.
3. AI encodes & magnifies bias.
4. Optimizing metrics above all else leads to negative outcomes.
5. There is no accountability for big tech companies.

At the end, I'll briefly share some positive ways that we can try to address these.

https://twitter.com/janellecshane/status/1405598023619649537

Challenges - What is the definition of fairness?

How to build models that are fair?



How would you prove that your model is not biased?

Types of Bias

1. Data Bias
2. Model Bias
3. Inference Bias

https://aws.amazon.com/sagemaker/clarify/

Data bias

If your data prefers a particular ethnic group or race, or age or accent, then the model trained on that data will also reflect or even amplify that bias

Similarly, a dataset that contains too many negative samples for one group may train a model to discriminate against that group

https://aws.amazon.com/sagemaker/clarify/

Model bias

A model can introduce bias if the prediction behavior is not consistent across different groups such as age, or gender, or income brackets

This behavior could be due to data or from bias introduced by the algorithm

"For instance, if an ML model is trained primarily on data from middle-aged individuals, it may be less accurate when making predictions involving younger and older people.“

https://aws.amazon.com/sagemaker/clarify/

Inference bias

The deployed model is showing signs of bias. The training data and the model were okay.

This can happen if the training data distribution and production data distribution are different

”For example, the outputs of a model for predicting home prices can become biased if the mortgage rates used to train the model differ from current, real-world mortgage rates”

https://aws.amazon.com/sagemaker/clarify/

SageMaker Tools

Detect bias and explain the model behavior

Clarify

Experiments

Model Monitor

Augmented AI

SageMaker Clarify

“Detect bias in ML models and understand model predictions”

- Unified capability (consolidates data from other tools)
- Detect bias
    - During data preparation
    - After model training
    - Deployed Models
- Tools to help explain model predictions

Explainability

Clarify uses model-agnostic feature-attribution approach to explain predictions

Uses game-theory to assign each feature an importance value [Shapley values]

- Why did the model reject a particular loan application?
- How does the model make predictions?
- Why did this model make an incorrect prediction?
- Which feature has the most significant influence on the behavior of the model?

Fairness metrics

How would you define fairness? How would you measure it?

Clarify Metrics

- At least eight different metrics for data bias
- Eleven other metrics for model bias
- Metrics to measure drift in live data

Metrics to quantify data bias

Measure Pretraining Bias - Amaz

docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html

aws | Search in this guide | English | Sign In to the

AWS > Documentation > Amazon SageMaker > Developer Guide | Feedback | Pre

Pretraining Bias Metrics

Bias metric	Description	Example question	Interpreting metric values
Class Imbalance (CI)	Measures the imbalance in the number of members between different facet values.	Could there be age-based biases due to not having enough data for the demographic outside a middle-aged facet?	Normalized range: [-1,+1] Interpretation: • Positive values indicate the facet *a* has more training samples in the dataset. • Values near zero indicate the facets are balanced in the number of training samples in the dataset.

Metrics to quantify post training model bias

Measure Posttraining Data and M

docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-post-training-bias.html

aws | Search in this guide | English | Sign In to the

AWS > Documentation > Amazon SageMaker > Developer Guide | Feedback | Pr

Posttraining Bias Metrics

posttraining bias metric	Description	Example question	Interpreting metric values
Difference in Positive Proportions in Predicted Labels (DPPL)	Measures the difference in the proportion of positive predictions between the favored facet *a* and the disfavored facet *d*.	Has there been an imbalance across demographic groups in the predicted positive outcomes that might indicate bias?	Range for normalized binary & multicategory facet labels: [-1,+1] Range for continuous labels: (-∞, +∞) Interpretation: • Positive values indicate that the favored facet *a* has a higher proportion of predicted positive outcomes. • Values near zero indicate a more equal proportion of predicted positive outcomes between facets. • Negative values indicate the disfavored facet *d* has a higher

Metrics to quantify drift in production model

Monitor Feature Attribution Drift

docs.aws.amazon.com/sagemaker/latest/dg/clarify-model-monitor-feature-attribution-drift.html

aws | Search in this guide | English | Sign In to the Console

AWS > Documentation > Amazon SageMaker > Developer Guide | Feedback | Preferences

- Monitor Model Quality
- Monitor Bias Drift
  - Create a Bias Drift Baseline
  - Schedule Bias Drift Monitoring Jobs
  - Inspect Reports for Data Bias Drift
- Monitor Feature Attribution Drift
  - Create a SHAP Baseline
  - Schedule Feature Attribute Drift Monitoring Jobs
  - Inspect Reports for Feature Attribute Drift
- Capture Data
- Schedule Monitoring Jobs
- Prebuilt Container

Monitor Feature Attribution Drift for Models in Production

PDF | Kindle | RSS

A drift in the distribution of live data for models in production can result in a corresponding drift in the feature attribution values, just as it could cause a drift in bias when monitoring bias metrics. Amazon SageMaker Clarify feature attribution monitoring helps data scientists and ML engineers monitor predictions for feature attribution drift on a regular basis. As the model is monitored, customers can view exportable reports and graphs detailing feature attributions in SageMaker Studio and configure alerts in Amazon CloudWatch to receive notifications if it is detected that the attribution values drift beyond a certain threshold.

To illustrate this with a specific situation, consider a hypothetical scenario for college admissions. Assume that we observe the following (aggregated) feature attribution values in the training data and in the live data:

College Admission Hypothetical Scenario

Feature	Attribution in training data	Attribution in live data
SAT score	0.70	0.10
GPA	0.50	0.20
Class rank	0.05	0.70

Complex collection of metrics!

"single, universal definition of fairness or a metric to measure it will probably never be possible. Instead, different metrics and standards will likely be required, depending on the use case and circumstances.“

Tackling bias in artificial intelligence (and in humans)

https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans

SageMaker Experiments

Need to optimize for predictive quality, and fairness of predictions!

May have to train 1000s of models

Hard to track best-performing models, and their input configurations

SageMaker Experiments

SageMaker Experiments automatically tracks the input, parameters, configurations, and results as trials

Clarify consolidates data to provide a feature importance graph to explain model's overall decision-making process after the model has been trained.

SageMaker Model Monitor

Continuously monitor quality of models in production

Configure alerts when deviations in model quality, bias drift

Model monitor is integrated with Clarify

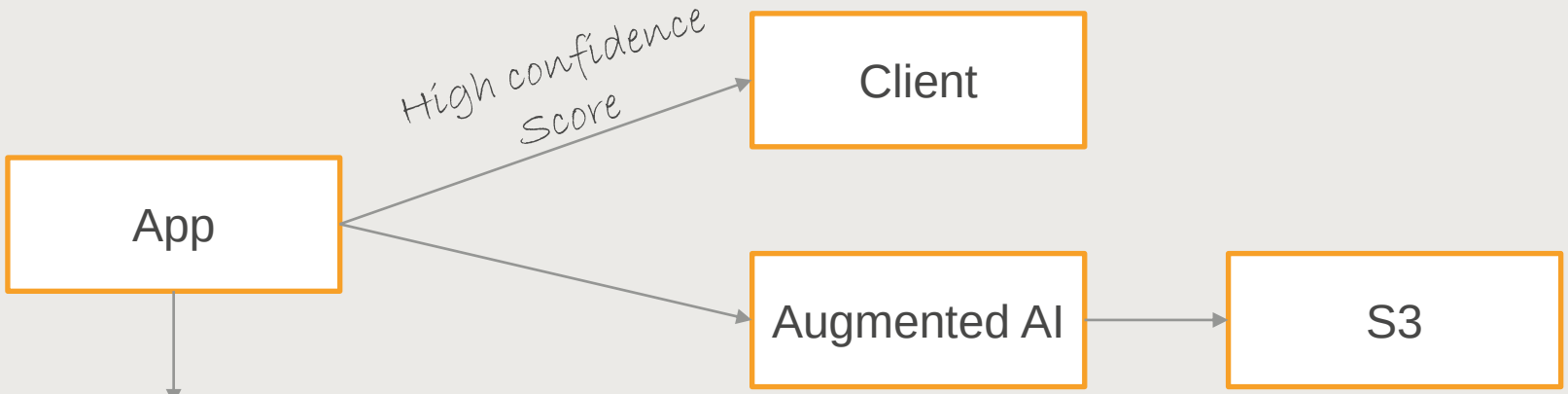
Amazon Augmented AI

Bring human-in-the-loop

Human oversight for ML predictions

Combine the benefits of ML and human-review

Amazon Augmented AI



Human Review

- Low confidence score
- Random sampling
- Single reviewer or multiple-reviewers

Augmented AI – Workforce options

Amazon Mechanical Turk	A crowd sourced marketplace of reviewers
	Suitable for public-data, non-confidential data
Private workforce	Reviewers are your own employees
	Ideal for customer confidential data
Labeling Service Providers (AWS Marketplace)	Suitable for customer confidential data
	Service agreement and clauses to protect customer data

SageMaker Tools

Detect bias and explain the model behavior

Clarify

Experiments

Model Monitor

Augmented AI

Summary

Regulatory Compliance

Policymakers, Regulators, Advocates

Ethics and policy challenges posed by AI

Companies may have to explain how AI makes decision

Internal Reporting and Compliance

Adoption of AI requires Trust

Explain behavior of trained models, How they make predictions

Customer Service

Financial advisors, Loan officers may review predictions made by AI system

Communicate to customers

AWS Certified Machine Learning Specialty

Chandra Lingam

70,000+ Students

Cloud Wave LLC