AI and Ethics (2021) 1:101–108
https://doi.org/10.1007/s43681-020-00006-3

OPINION PAPER

Using unethical data to build a more ethical world

How CallMiner handles imperfections in speech recognition

Jamie Brandon[1]

Received: 27 August 2020 / Accepted: 29 August 2020 / Published online: 6 October 2020

Abstract
Data scientists use data to train models. Those models calculate probabilities to capture patterns in the data. It's difficult to build ethical models when the available training data contains racism, sexism, or other stereotypes. Contact center data, including calls, chats, texts, and emails, is no exception. Instead of building a model to automate decision-making processes, we use the unethical findings from our model as an insight. We discuss debiasing options for removing racism from the model but find that removing this bias removes a crucial insight that an analyst deserves to know. By leaving the model with all the biases learned from the training data, we can provide better analytics. Analysts can recommend solutions that start to dismantle the systemic racism present in our society. Debiasing is not always appropriate. Censoring the model makes it harder to identify what can be done to prevent racism in our procedures and society.

Keywords Ethics · NLP · Word embeddings · Debiasing

1 Introduction

When a model performs poorly, it's easy to blame the data. After all, the model simply captures patterns from the training data, like quick restaurant service correlating with a positive review. That model might be used to predict whether a new, unlabeled review is positive or negative. Models can also be used descriptively, showing insights to what might be causing the positive or negative reviews. When poor performance occurs, someone might blame the data. Maybe there's not enough data to differentiate between positive and negative. Maybe there should be a category for neutral sentiment too. Perhaps data instances were labeled incorrectly, skewing the classifier in the wrong direction. There are plenty of ways the data can be incomplete, inconsistent, or inaccurate. Dirty data affects model performance, but model performance should not be the sole indicator of success.

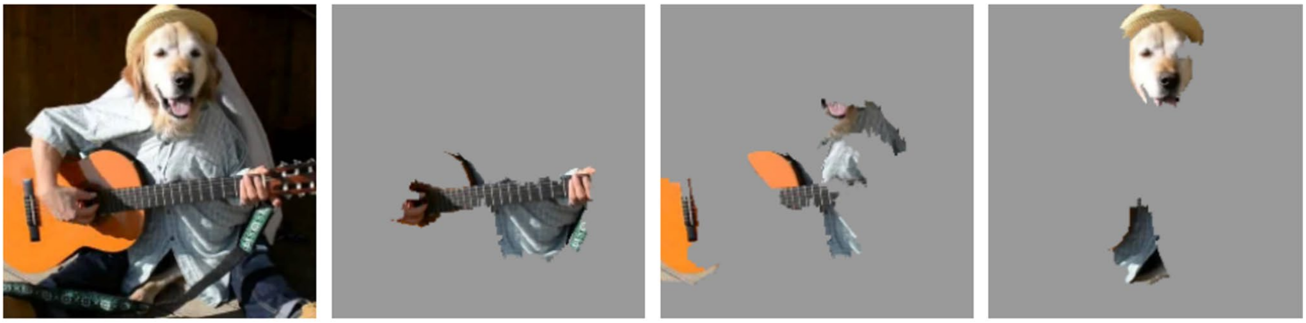
A model with a 20% accuracy score should not be put into production. A model that makes racist decisions 80% of the time should not either. That is to say, models and data can be ethically dirty too. The model trained on unethical data may carry harmful notions about race, gender, etc. despite performing well on a test set. When a model captures the unethical bias from the training data, it's easy to accidentally perpetuate harmful stereotypes. As practitioners, we can do more to protect marginalized groups. It's not enough to simply blame the data for an unethical model.

Data scientists usually carry no intention of building an unethical model. The bias exists in the training data, so the model captures that pattern. For example, when selecting features to build a model, someone might include zip code. They know that a person's residence has an influence, but they do not realize that zip code correlates strongly with race. By including zip code as a feature, they have accidentally built a model that uses information about race when predicting. Features that highly correlate with race or gender pose ethical dilemmas when modeling.

Unrepresentative data sets pose another ethical problem. Consider a computer vision task: given a photo containing a bride, the model's goal is to place a bounding box around her. The training data includes photos from beach weddings and chapel weddings. The model captures the features that distinguish a bride in these photos: a white dress, a bouquet of flowers, maybe a veil. The model excels at this specific task, but brides have a wider range than what the training

✉ Jamie Brandon
jamie.brandon@callminer.com

[1] CallMiner, Waltham, USA



(a) Original Image **(b)** Explaining *Electric guitar* **(c)** Explaining *Acoustic guitar* **(d)** Explaining *Labrador*

Fig. 1 Explaining an image classification prediction made by Google's Inception and neutral work. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$). Reprinted from "Why Should I Trust You?" Explaining the Predictions of Any Classifier, as cited in [1].

data includes. A bride in India might wear bright red instead of white. What about a wedding photo with two brides? Will the model be able to place a second bounding box around a bride's new wife too? If none of the training samples contain brides like these, the model will not perform well on all brides. Representative training data sets are crucial to ethical model building.

Models are really good at predicting the future when the future looks exactly like the past. Since our past contains systemic problems with human rights and siloed approaches to collecting training data, we need to be aware of accidentally propagating harmful stereotypes. What can a data scientist do when presented a data set riddled with unethical biases? Annotators could remove all data instances that contain unethical bias, but they might not realize when a data instance carries an unethical notion about an unfamiliar culture. It would take a lot of time and resources to comb through a data set. At prediction time, a model in production may be presented with new instances that contain unethical data. It would not perform well on these instances if it did not have any like them in the training set. Removing the unethical instances would not solve the problem. We need a way to make sure the model is learning the right lessons from unethical data.

Instead of using models for prediction engines that automate decision making, we can use unethical models as insight engines. A model that outputs racist predictions exposes the racism in the training data, and by extension, our society. Knowing that these problems exist on a large scale provides leverage to marginalized communities. The biased model shows evidence for the bias in our world. With humans learning about how and why these problems occur, we have a better avenue for instigating change.

I'll discuss some examples of models that propagate stereotypes. I'll talk about my experience as a data scientist working for a company offering speech and text analytics to contact centers. While it's difficult to build an ethical model from unethical training data, the insights gained by modeling are often more crucial than the prediction. We can use these insights to instigate change in our communities, our society, and our policies.

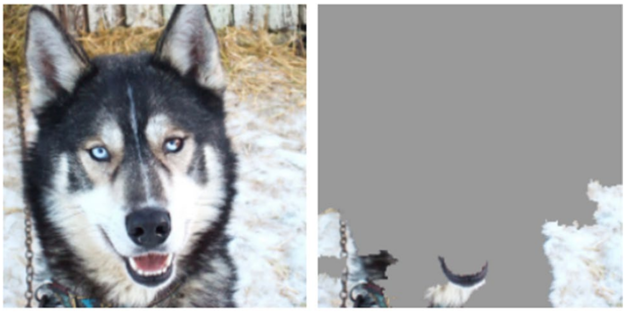
2 Unintentionally unethical models

It's difficult to build an ethical model given training data from unethical sources. Models built from traditional machine learning algorithms don't take causation into account, only correlation. If a feature correlates strongly enough with an outcome, the model places heavy weight on that feature, whether it's something truly indicative of the outcome or not.

In the paper *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*, the authors show that the models look for strong correlations to achieve better model performance [1]. They provide a lens into the model's decision-making process by shadowing portions of the image that do not play a large role in determining the label. See the image below to see what the model attends to when assigning the labels for "Electric Guitar," "Acoustic Guitar," and "Labrador." (Fig. 1)

In another task, the authors build a classifier to distinguish wolves from huskies [1]. The authors intentionally select twenty photos to use in their training set. All pictures with wolves include snow, and all pictures with huskies do not. When given a test image of a husky with snow (below), the model attends much more to the snow in the picture than to the subject. It erroneously classifies a dog as a wolf (Fig. 2).

Models will attend to spurious correlations. There is not a way to tell the model that snow is irrelevant when determining wolf or husky. A model has no world knowledge about huskies, wolves, or snow. It does not even have the question "What's the difference between a wolf and a husky?" All it



(a) Husky classified as wolf **(b)** Explanation

Fig. 2 Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task. Reprinted from "Why Should I Trust You?" Explaining the Predictions of Any Classifier, as cited in [1].

has are two sets of pictures and an algorithm to find features that distinguish the two sets. What a model uses to predict is not always what the data scientist intends.

Applying this concept to human data shows the challenge in ethical modeling. Models will use correlations caused by racism, sexism, and other flaws in society. Given training sets like historical court cases [2], a model would predict that Black criminals should be sentenced for longer jail times than white criminals with the same offense. Given training sets like medical data in which women's pain is not considered seriously [3], women with chronic conditions remain undiagnosed and untreated.

Models also have to deal with nefarious users when accepting unfiltered training data. Take for example Microsoft's Twitter Bot, Tay, from 2016. The bot was designed to interact with the users of Twitter and learn from each interaction. Within the first several hours, users of Twitter started intentionally feeding the bot data with politically incorrect statements. The bot has no moral compass. The bot knows only conditional probabilities from the data it's been trained on. Because it's been fed hurtful language, it's likely to repeat hurtful language. Microsoft took the bot down after less than 24 hours of live interactions [4]. They were unable to quickly find a way to keep their bot ethical after being trained on unethical data.

There are plenty of other examples of scientists creating racist models without ill intentions. Robyn Speer even offers a tutorial, *How to make a racist AI without really trying* [5]. The real-world data sets we use for modeling have the same biases our society does. It's not surprising that the model captures that bias. When a model has all the bias that our world has, it's extremely likely to stifle progress for marginalized groups and perpetuate the privilege of advantaged groups. We need to ensure that marginalized groups don't continue to be marginalized because of a model's outputs.

Table 1 Error rates on matched subset of identical short phrases spoken by white and Black individuals in our sample. Adapted from "Racial disparities in automated speech recognition", as cited in [6].

	Average WER for Black speakers	Average WER for white speakers
Apple	0.28	0.12
IBM	0.21	0.10
Google	0.17	0.11
Amazon	0.18	0.08
Microsoft	0.13	0.07

3 Bias in speech recognition

Unintentionally unethical models are prevalent, and speech recognition is no exception. Speech recognition is notoriously difficult across diverse voices, regions, ethnicities, and dialects. Given a matched subset of identical short phrases spoken by Black and white speakers,[1] the word error rate (WER) for Black speakers is higher than white speakers [6] (Table 1).

When the majority of data used to train the model comes from a limited demographic of humans, a model will struggle to perform well on unrepresentative data samples. It's likely that when acquiring training data, diversity metrics weren't considered. Being neglectful of diversity yields technology that does not work as well for certain groups of people. In this case, it is likely Black speakers were not accurately represented, and any resultant model will contain this bias. In another article, a Black speaker admits to quitting his use of Siri because of its poor performance. "Having to adapt our way of speaking to interact with speech technologies is a familiar experience…" the author states [7]. Making technology that only works well for white people is asking others to assimilate to a white voice, a white mindset, a white culture.

There is a vast diversity of accents and dialects from customers and contact center agents around the world. It's important to train models from a broad sample of voices. If a speech recognition model has been trained only on young white male voices from the Midwest, the model's performance on an elderly woman from England will suffer.

[1] I like the perspective *The Seattle Times* offers in their style guide.

Black (adj.): Belonging to people who are part of the African diaspora. Capitalize Black because it is a reflection of shared cultures and experiences (foods, languages, music, religious traditions, etc.) …

white (adj.): Belonging to people with light-colored skin, especially those of European descent. Unlike Black, it is lowercase, as its use is a physical description of people whose backgrounds may spring from many different cultures.

Training a model to capture every unique dialect and accent is extremely difficult.

Speech recognition is often the first step in a pipeline. Virtual assistants will use speech recognition as an input to an intent classifier. Given a user's speech, *"What's the weather?"* the automatic speech recognizer transcribes *[What's a weather?]*. The intent classifier takes *[What's a weather?]* and determines that this is a query indeed for weather, despite the small error from the recognizer. When the error in step one is larger, it affects downstream tasks significantly. For use cases that depend on speech recognition, it's important to take into account that the data from step one contains unethical bias.

4 The effect of dirty data on contact center analytics

As a data scientist at CallMiner, I am responsible for handling dirty data and any unethical bias that accompanies it. CallMiner's product provides speech and text analytics to contact centers. I work with the research team to extract insights from call center conversations, chats, emails and more. Some contact centers are looking to increase their positive customer experience; others want to increase their sales effectiveness. Still others are a bit overwhelmed with the sheer amount of data they have. Their question is vague: "What don't I know about my call center or my customers?" Our product is built to be flexible. Users can search for actionable insights. These insights drive better business decision making, allowing businesses to learn from past contact center interactions.

It's difficult to do proper analytics work when the data to be analyzed is dirty. The analysis relies on the data collected from our clients, mainly over-the-phone conversations. Some of the dirt in our data comes in the form of incorrect speech-to-text transcription. Callers and agents come from many different cultures and locations. It's difficult to capture this wide range of diversity in a single speech recognition model. The data scientists on our research team receive this noisy data, and we take extra precaution to handle the errors.

Our product allows for automated agent scoring. Instead of being judged by humans on only a handful of their interactions, they are now scored consistently on every interaction with a customer. This guarantees that each agent is scored on the same criteria regardless of whether their manager favors them. However, automation poses ethical challenges. If the speech recognizer struggles with an agent's accent, they may not be scored the same. Take for example agent Alice and agent Bob. Bob answers all his calls with the company's recommended introduction, "I appreciate you waiting in the queue." Alice has a speech impediment. Even though she says the same thing, the transcription reads, "I pizza ate you wading in the pool." Her transcription is close to meaningless when read as is, but the caller responds normally, showing that they understand her without a problem.

When an analyst searches for the recommended greeting, they search for phrases like *[I appreciate]* and *[waiting in the queue]*. Bob's calls are found, but Alice's aren't. Will Alice's leader think she's performing worse than Bob? What about when analysts start looking for legal compliance language? Some agents are legally required to recite a short script. If a recognizer captures a phonetic variation of that script, it's difficult for the analyst to know whether the call followed the legal protocol. How can we make finding these consistent speech misrecognition errors easier? If the goal is to predict customer satisfaction from the transcript, I want to ensure that both Alice and Bob's introductions are treated as similar inputs to the model. When Alice is compliant, it's unfair for Alice to receive a lower score because a speech recognizer doesn't capture her words properly.

The data we're using as input to predictive models already contains the biases reflected in society. When a model is trained on that biased data, the model will hold the same biases represented in the data set. How can we account for errors in speech transcription before giving this data to a predictive model? How can we help our clients make more robust searches? How can data scientists ethically predict an outcome if the input is riddled with unethical biases?

5 CallMiner's solution: Illuminate

CallMiner is aware of the disparities in speech recognition. We realize that speech recognition is a difficult task, so there will not be perfect results for every speaker. We advise analysts to search for phonetic variations of their intended query. When trying to find *[enunciate your syllables]*, we encourage searching for *[in Nancy eight your syllabus]* too. We call these phonetic variations aliases. Coming up with these aliases can be difficult though. How does an analyst know how something will get misrecognized?

An analyst would benefit from a tool that recommends relevant aliases based on historical data. To build this, CallMiner's research team trains a model that captures the similarity between aliases and their intended phrase. The model knows that *[have a nice day]* and *[have an ice day]* are very similar to one another. This similarity does not depend on the definition of the word. What matters is the context that word fell in. Words that fall into similar contexts have similar meanings. This is the distributional hypothesis, and it holds over the noisy channel of speech transcription.

Think about the examples below.

Intended statement	Transcription with error
I need to feed my **cat**	I need to feed my **CAD**
I need to feed my **dog**	I need to feed my **bog**
I forgot to feed my **fish**	I need to feed my **wish**
I need to feed my **baby**	I need to feed my **lady**

All of the bolded words fall into very similar contexts, namely preceded by *[to feed my]*. The first three examples capture the specific concept of *[pet]*, while the last one expands this concept to *[dependent living beings]*. Even if a speech recognizer erroneously transcribes these words, the contexts are similar enough that *[cat]* will be close to *[CAD]* and *[dog]* to *[bog]*. If a word is consistently misrecognized, we can find that alias by finding the most similar words to the intended word.

We trained our own word embedding model [8] on call center transcriptions (and call center transcriptions only). We can then define similarity between words to be the similarity between word vectors. We wanted to let the data tell us what the words mean, not begin with a pre-trained model that already knows that *[canine]* and *[dog]* are similar. What if, more often than not, those are aliases in call center data?

We trained each client's model separately from every other client. While this does mean each model is trained on less data, that data is representative of the call center in consideration. For example, in one client's data set, *[lemon]* is an alias for *[limit]*; while in another, it's referring to lemon-scented wood polish. By training a model for each client, that alias does not get polluted with contexts that aren't relevant. Common misrecognition from other clients don't influence the word embeddings for everyone, only their specific model.

Our product, Illuminate, aids users who are searching through their calls. Before Illuminate, when an analyst was searching for phrases like *[There's an issue]* or *[I'm having trouble]*, they would have to think of possible aliases on their own. Maybe a speech recognizer would erroneously capture *[There's my tissue]* or *[I'm in a bubble]*. All of these guesses for possible aliases need to be thought up and tested by a human.[2] With Illuminate, words that consistently alias are similar to the intended word. So instead of thinking "what's an alias for *[issue]*?", an analyst can search for words most similar to *[issue]* in their own custom word embedding model. If *[tissue]* is a consistent misrecognition, the model will show *[tissue]* as similar to *[issue]*, and the analyst can feel confident adding it to their search. These expanded searches are more robust. They capture the perfect transcriptions as well as some of the imperfect transcriptions. Despite the imperfections from the speech recognizer, people with accents can be scored without having to assimilate their speech patterns.

This is a great first step toward a more ethical searching in our product. Users can now easily make searches that include aliases without being burdened by the mental load of "How would a speech recognizer get this a little wrong?" We are able to take a biased input and use it for a more ethical outcome.

6 Do word embeddings capture bias?

All of this sounds like a great solution to handling unethical inputs for modeling, but this is not a bias-free solution. As the statistician George Box famously said, "All models are wrong, but some are useful." Even though this model makes for a more ethical usage of our product, the model was built from data that contains bias. It'll be biased too.

Because the models were trained on contact center interactions, they carry all the biases present in that data. When searching for words similar to *[idiot]*, racial slurs are often returned. These words are mathematically similar to one another because they occur in similar contexts in the training data. While a contact center can work to control the way agents talk about race, there's little they can do to control what a customer says. Since all of this unwieldy customer data is given to the model as training data, the output of that model will capture the racial bias.

Despite our goal to overcome the imperfections in speech recognition, our model captures the racism present in client interactions. To combat this, we've released a disclosure to analysts working in the product. This disclosure has examples of how to interpret a racist result from the model and an explanation for why this result is being returned. When the model returns a slur after searching for *[stupid]*, it doesn't mean anything about the intelligence of that group of people. It means that in the contact center data, people are using both of these terms in similar contexts. It would be inappropriate to include this term in the search for a topic regarding agent knowledge, but it would be appropriate to add it to a search for a topic regarding insults. It's not the best practice to blindly include every term the model suggests. We still require a human in the loop for building ethical searches.

With the warning of bigoted results, analysts are able to play an active role in changing their call center, and by extension, their society.

[2] CallMiner even offers a product, SearchQA, that allows analysts to check the search they've created. This allows a human to add a final layer of sanity to their analytics.

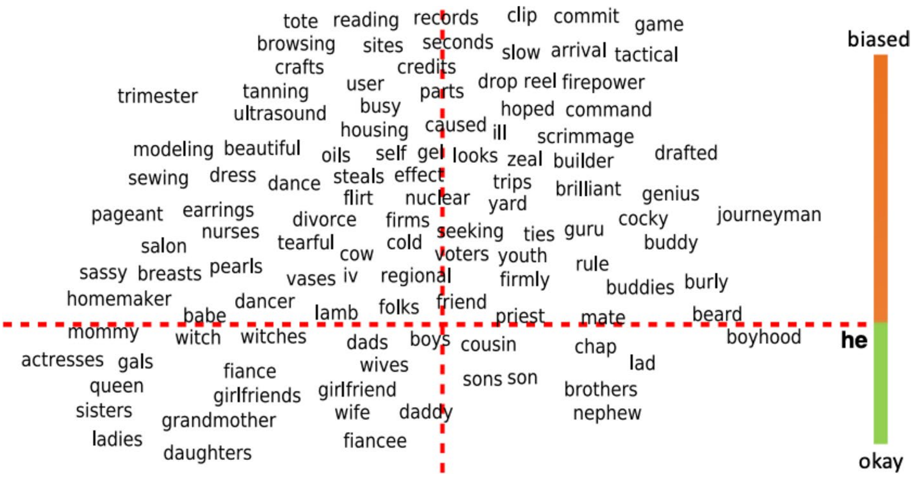


Fig. 3 Selected words projected along two axes: x is a projection onto the difference between the embeddings of the words he and she, and y is a direction learned in the embedding that captures gender neutrality, with gender neutral words above the line and gender specific words below the line. Our hard debiasing algorithm removes the gender pair associations for gender neutral words. In this figure, the words above the horizontal line would all be collapsed to the vertical line. Reprinted from "Man is to Computer Programmer as Woman is to Home-Maker? Debiasing Word Embeddings" as cited in [8]

7 Intentionally (more) ethical models

There are papers about decreasing the bias in word embedding models. Notice the intentional word choice: *decrease* instead of *eradicate*. While it's possible to take steps to control the bias we know exists, it's much harder to control bias we are not aware of. A model will always be biased, but we can take steps to mitigate the unethical bias.

Bolukbasi et al. in the paper *Man is to Computer Programmer as Woman is to Home-Maker? Debiasing Word Embeddings* discuss the task of the SAT-style analogy questions and the role biased word embeddings play. In this task the model is presented with an analogy question like "Man is to computer programmer as woman is to ________." The model predicts the word to fill in the blank using word embeddings. The authors propose and discuss a two-step process to debias word embeddings: identifying the gender subspace and neutralizing and equalizing [8]. The gender subspace can be visualized below. The x-axis captures notions about gender, with words like *[mommy]*, *[queen],* and *[daughters]* on the left and *[brothers], [sons],* and *[nephew]* on the right. The y-axis indicates how biased these terms are. Words that have nothing to do with gender, like *[sewing], [reading],* and *[dancer]* all fall on the female side. On the upper right, we see words like *[brilliant], [cocky],* and *[builder]* (Fig. 3).

Neutralizing guarantees that gender-neutral words will be 0 in the gender subspace. Equalizing ensures that word sets like *[sister]* or *[brother]* are equidistant from the gender-neutral counterpart *[sibling]*. Doing this alters the word embeddings so that the gender of a word does not influence the meaning in cases where it should not, like professions. In cases where gender is encoded into the meaning of the word, like *[mother]* or *[father]*, then that meaning is still captured by the word embeddings.

The results of the paper are promising. In the analogy prompt *[he is to doctor as she is to X]*, the original model returns *[nurse]*. The debiased model returns *[physician]*. It also maintains relationships like *[he is to prostate cancer as she is to ovarian cancer]*. In medical research, linking the strong correlation of gender to anatomy may prove useful in modeling. However, the word embeddings capture a notion that anatomy indicates gender, which is harmful to the transgender community.

This paper fails to address gender as more than binary. While their work on debiasing is profound, it's disappointing that the full potential of the problem was not addressed. I do not expect one paper to solve all the problems at once but be sure to acknowledge that the solution is incomplete. As scientists, policymakers, and members of our communities, we need make consideration for whomever might be excluded when posing a new solution.

This paper offers one solution for debiasing word embeddings. For analogy problems, results are updated so that sexist notions in society are permeated less strongly. While the work is promising, it's important to consider that some folks were left out of this solution. Despite the flaw, this solution moves the field in more ethical direction.

8 Does debiasing solve the problem?

CallMiner's specific challenge is different from the analogy question posed above. We're building word embeddings tailored to each client's data. We capture unique aliases that each client has. We're also capturing the unique ethical challenges each client faces.

After discussing the options available for debiasing word embeddings in each client's model, we made the decision to leave the model with all the biases present. An analyst deserves to know when their agents are being verbally abused by customers. They also deserve to know the reason, whether it's gender, ethnicity, accent, race, or simple human frustration. Censoring these unethical results blinds the analysts and companies to the reality their agents face every day. The whole point of the tool was for the analyst to be able to search their interactions more robustly. Censoring content makes an analyst's job harder.

Instead of debiasing, we choose the approach of educating the product user about ethics in machine learning. Let's consider the example of racial slurs being suggested when an analyst searches for *[stupid]*. By teaching an analyst that the model returns racial slurs because they're used in similar contexts as [*stupid]*, the analyst gains awareness of the hardships an agent faces every day. This offers an actionable insight. The call center might need better training for how to handle racist interactions. The agents might need leadership to intervene in interactions where a customer accosts an agent. The analyst could recommend that agents take a break after difficult interactions, decreasing the company's rate of agent turnover. We've empowered the analyst by using the biased data from their company's interactions as an insight engine.

The researchers at CallMiner know that this solution is specific to our use case. We are building models based on data from one client at a time. For an agglomeration of data, debiasing seems more appropriate. When a human is not in the loop, debiasing seems more appropriate. When we're using the model as an insight engine, debiasing removes valuable insights the analyst could use.

Reporting the unethical results can help to diminish the propagation of unethical stereotypes. The American Association for University Women published their analysis on how transparent salary information decreases the wage gap. There is a 13% pay gap between men and women in federal jobs where salary information is public. In state governments, where salary information is often published, the pay gap is 18%. In the private, for-profit sector, where there is little transparency, the pay gap is 29% [9]. As the transparency decreases, the pay gap increases. By publishing the insightful results from a model, an analyst becomes a catalyst for change. Being transparent is crucial to building a more ethical society.

Debiasing a model censors out insights around unethical bias. Analysts need to be able to find these insights and share them with management, peers, and their society. Evidence shows transparency with biases instigates change. By choosing not to debias the model, the analyst will be more aware of problems that exist. Just as reporting the wage gap in salaries, the model showing its flaws can lead to solutions in society rather than just in the model. Before finding a solution, awareness and a proper understanding of the problem are key.

9 Conclusion

The manner in which someone intends to use a model should guide the data scientist about handling ethical biases. We know about the racial tensions present in our society, so it's no surprise that we find those racial tensions in contact center data too. Leveraging the model's unethical results as a way to bring awareness to an issue might be more painful for an analyst to see every day, but the agents are feeling it every day too. Blinding the analyst to the issue at hand doesn't solve the problem. By using word embeddings as a word similarity tool, analysts can extract better insights about the problems their company faces. It also helps them build more robust searches to evaluate agents regardless of their speaking patterns.

As data scientists and stakeholders, we must be cognizant of our products' impact. Contact center workers come from different cultures. Many have accents and speak more than one language. Training a speech recognizer on the vast differences may be difficult, but don't let "garbage in" be an excuse to output garbage as well. Debias where you can. Educate users when censoring out the bias will blind the user to an important insight.

When we developed Illuminate, we discovered the ethical ramifications after building the model. While disappointing to find, these results validate the experience that marginalized populations have been experiencing for centuries. We hope our product allows analysts a better way to extract insights. We hope their searches find more than one way to say something. We hope other data science teams follow our lead: make the right solutions to instigate change.

Acknowledgements With acknowledgements to Rick Britt, Chris Jones, Anya Korneyeva, Micaela Kaplan, Kirsten Stallings, Bill Muller, Larry Medsker, and Keegan Parrotte for their reviews

References

1. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 2016.
2. Rehavi, M.M., Starr, S.B.: Racial disparity in federal criminal sentences. J. Pol. Econ. **6**, 1320–1354 (2014)
3. Kiesel, L.: Women and Pain: Disparities in Experience and Treatment. Harvard Health Publishing, Boston (2017)
4. Miller, K.W., Wolf, M.J., Grodzinsky, F.S.: Why we should have seen that coming: comments on microsoft's tay "experiment", and wider implications. Orbit J. **1**, 1–12 (2017)
5. Speer, R.: How to make a racist AI without really trying, ConceptNet Blog, 13 July 2017. http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/. Accessed 8 Sep 2020.
6. Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J.R., Jurafsky, D., Goel, S.: Racial disparities in automated speech recognition. PNAS **117**(14), 7 (2020)
7. Lloreda, C.L.: Speech Recognition Tech Is Yet Another Example of Bias, Scientific American, 5 July 2020. https://www.scientificamerican.com/article/speech-recognition-tech-is-yet-another-example-of-bias/. Accessed 8 Sep 2020.
8. Bolukbasi, T., Chang K.-W., Zou J., Saligrama V., Kalai, A.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: 30th Conference on Neural Information Processing Systems, 2016
9. American Association of University Women, "Salary Transparency Linked to Smaller Gender Pay Gap," 19 November 2019. [Online]. Available: https://www.aauw.org/resources/news/media/press-releases/salary-transparency-linked-to-smaller-gender-pay-gap/. Accessed 8 Sep 2020.