DATA FEEDBACK LOOPS: MODEL-DRIVEN AMPLIFICATION OF DATASET BIASES

Anonymous authors
Paper under double-blind review

ABSTRACT

Datasets scraped from the internet have been critical to large-scale machine learning. Yet, its success puts the utility of future internet-derived datasets at potential risk, as model outputs begin to replace human annotations as a source of supervision. In this work, we formalize a system where interactions with one model are recorded as history and scraped as training data in the future. We then analyze its stability over time by tracking changes to a test-time bias statistic (e.g. gender bias of model predictions). We find that the degree of bias amplification is closely linked to whether the model's outputs behave like samples from the training distribution, a behavior which we characterize and define as consistent calibration. Experiments in three conditional prediction scenarios – image classification, visual role-labeling, and language generation – demonstrate that models that exhibit a sampling-like behavior are more calibrated and thus more stable. Based on this insight, we propose an intervention to help calibrate and stabilize unstable feedback systems.

1 INTRODUCTION

Due to the successes of large-scale training in machine learning (He et al., 2016; Brown et al., 2020; Radford et al., 2021), datasets derived from publicly available internet data have become indispensable to the machine learning community. For example, without relying on internet scraping, it would be cost-prohibitive to manually construct key datasets such as ImageNet (Deng et al., 2009), The Pile (Gao et al., 2020), or YFCC100M (Thomee et al., 2016). While the internet has served as a large, easily-accessible source of human generated data in the past, the growing deployment of machine learning systems puts this procedure at risk. As models begin to create and annotate a significant fraction of internet content, the utility of the internet as a data source may decrease rapidly.

As an example in visual role-labeling, consider a classifier trained on public photos and their associated tags, as depicted in Figure 1. Instead of manually tagging photos, some users may instead choose to auto-tag their photos with the model. These photos, now stored in internet history, may be scraped as training data for an updated iteration of the image-tagging model. Any systematic biases introduced by the model, such as consistently mislabeling female doctors as nurses as in Figure 1, are now encoded into the training data. This *data feedback* gradually degrades the quality of the internet as a data source, since supervision becomes driven by model outputs rather than human annotation.

Issues stemming from having previously model-generated content included in training data have already been encountered in machine translation (Venugopal et al., 2011) and speech recognition (Radford et al., 2022). These concerns are especially important in situations where model predictions may exacerbate existing toxicity, harm, or other biases (Gehman et al., 2020; Zhao et al., 2017). In such cases, a viable strategy for model developers is to weigh the benefit of updating their model to new internet content versus the cost of amplifying biases via such model-induced feedback. However, it is not yet understood when and to what degree data feedback is an issue in practice.

In this work, we define the data feedback setting and carefully study how model biases change under feedback. In particular, we ask: Are there conditions that stabilize bias amplification? We answer this in the affirmative, finding that one crucial path to achieving stability guarantees is having a consistently calibrated training procedure – one that produces models with a bias similar to its training distribution. Furthermore, this form of calibration can be realistically achieved in natural experimental settings. Specifically, models that behave like samplers (i.e. replicate their training distribution well) are more likely to be calibrated and thus more stable. In addition, many prediction algorithms that do not explicitly perform sampling, such as image classifiers, fulfill this behavior through a conjectured phenomenon called Distributional Generalization (Nakkiran & Bansal, 2020).

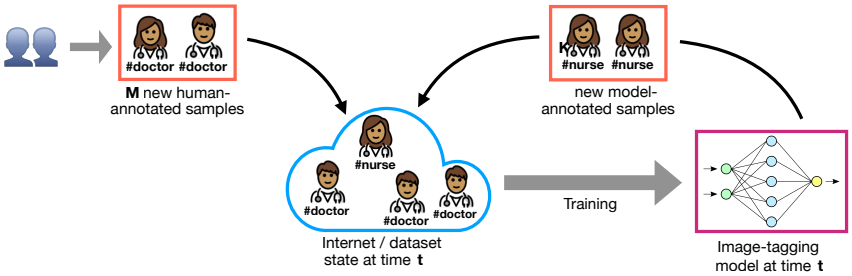

Figure 1: A simple example illustrating data feedback. An image-tagging model is trained on images from the internet. Some users choose to auto-tag new images with the model and post them online, while others continue to manually tag their images. After some time, the model may be updated by re-scraping the internet and re-training on the updated dataset, which now includes feedback from previous model predictions.

Formally, we quantify the stability of data feedback with a bias metric $\phi(x, \hat{y})$, where $\hat{y} = f_t(x)$ are predictions from the model at time t. For example, the predictions $\hat{y}$ are image tags or sentence completions, and the bias metrics ϕ are gender bias or sentence toxicity. Our theoretical result shows that if the model does not increase bias by more than error δ, then the total bias amplification is bounded by $\frac{m+k}{m}\delta$, where m and k refer to the number of new human-annotated samples and model-annotated samples respectively. Thus both a smaller calibration error δ and a higher fraction of human-annotated samples m contribute to the global stability of data feedback loops.

The rest of the paper is organized as follows. In Section 3, we define the data feedback setting in more detail. We then describe a specific notion of calibration (consistent calibration), discuss its connection to sampling, and show how it gives rise to bounds on bias amplification in Section 4. Section 5 demonstrates the utility of these predictions empirically in three different natural experiment settings:

1. First, we define a simple data feedback setting in CIFAR (Krizhevsky, 2009), where the label distribution is skewed and feedback has the potential to amplify label shift. In this case, we show the feedback dynamics are stable and consistent with our theoretical predictions.
2. Next, we show that data feedback can significantly amplify gender biases in a visual semantic role labeling task (Yatskar et al., 2016). Our bounds predict that the dynamics may be unstable since the initial calibration error is large, which is consistent with gender bias amplification identified in earlier work (Zhao et al., 2017).
3. Third, we examine data feedback for language generation on a toxic prompts dataset (Gehman et al., 2020) and demonstrate that toxicity and repetition amplify, with sampling-based generation schemes enjoying substantially higher stability than beam search methods.

Finally, to conclude Section 5, we design an intervention to stabilize beam search methods by leveraging the sampling-like behavior of interpolating classifiers (Nakkiran & Bansal, 2020). To do this, we train a language model that overfits to its training set and observe that this procedure significantly stabilizes the model's toxicity and repetition.

2 Related Work

Performative prediction. The general problem of model-induced feedback in machine learning has been previously studied as performative prediction and strategic classification (Perdomo et al., 2020; Hardt et al., 2016), where future data distributions can change arbitrarily in response to the deployed model. In this context, existing work has focused on methods that optimize towards equilibria of the system (Brown et al., 2022). The generality of the problem setting allows for complex human interactions in-the-loop; however, it is for this reason that experimental evaluation has been limited, and most analyses have focused on convex settings with experiments on Gaussian data or simple synthetic data such as loan applications or credit risk (Izzo et al., 2021; Miller et al., 2021).

In contrast, motivated by the image tagging example in Section 1, we consider a more restricted form of feedback, in which new data examples are gathered only from either the "true" human-annotated distribution or predictions of the currently deployed model. This restriction allows us to analyze feedback stability in more realistic experimental settings and derive bounds on stability.

Bias amplification. Machine learning models have a tendency to amplify at test-time biases that exist in their training data, a problem known as bias amplification (Dinan et al., 2019; Leino et al.,

Algorithm 1 Data Feedback Procedure

Input: Human-annotated distribution P_0, training algorithm $\mathcal{A}$, initial number of samples n_0, human-annotated samples per round m, and model-annotated samples per round k
Output: Model deployments over time $f_0, f_1, f_2, \ldots$
1: $\mathcal{S}_0 \sim \mathsf{P}_0^{n_0}$
2: Deploy $f_0 \sim \mathcal{A}(\mathcal{S}_0)$
3: **for** $t \in \{1, \ldots \infty\}$ **do**
4: $\quad \mathcal{S}_t = \mathcal{S}_{t-1} \cup \{(x_i, y_i)\}_{i \in [m]} \cup \{(x_j, f_{t-1}(x_j)\}_{j \in [k]}$, where $(x_i, y_i) \overset{\text{iid}}{\sim} \mathsf{P}_0$ and $x_j \overset{\text{iid}}{\sim} \mathsf{P}_0(x)$.
5: $\quad$ Deploy $f_t \sim \mathcal{A}(\mathcal{S}_t)$
6: **end for**

2019; Hall et al., 2022). For example, image classifiers have skewed gender predictions, beyond what exists in the training data (Zhao et al., 2017; Wang et al., 2019). In our work, we build on this literature by studying the multi-step amplification of bias via feedback.

Feedback in healthcare. The data feedback setting is most related to feedback loops previously studied in healthcare (Adam et al., 2022; 2020), where false positive examples are added into the training set over time. These works have focused on benchmarking prevention and mitigiation strategies in the context of binary classification on tabular data. Understanding how well these methods work in more general settings is important future work.

Additional discussion relating to recommender systems, semi-supervised learning, domain adaptation, and more can be found in Appendix A.

3 Defining Data Feedback and Model Bias

Our work considers feedback effects in the conditional prediction setting. In the standard conditional prediction or supervised learning framework, the goal is to learn a function $f \in \mathcal{F}, f : \mathcal{X} \to \mathcal{Y}$ from a collection of samples $\{(x_i, y_i)\} \overset{\text{iid}}{\sim} \mathsf{P}_0$. P_0 represents a fixed human-annotated example distribution (e.g. human-tagged images or human-written prompts and completions). Motivated by Figure 1 where the dataset changes over time, we instead consider a series of supervised learning problems from time $t = 0 \ldots \infty$. At each time, we learn a new model f_t using the latest available internet data.

The series of supervised learning problems are defined by the following. At $t = 0$, before any data feedback, only clean human-annotated samples are available on the internet. Thus, the initial model f_0 is trained on n_0 i.i.d. samples from P_0, and we call this initial dataset $\mathcal{S}_0 \sim \mathsf{P}_0^{n_0}$ and the resulting model $f_0 \sim \mathcal{A}(\mathcal{S}_0)$. Here, $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{F}$ refers to a potentially stochastic learning algorithm, which we take to be a neural network trained on the cross entropy loss with SGD.

For any $t \geq 1$, we assume that data on the internet grows in two ways. Humans naturally continue to interact with the internet and generate data, creating m new samples following the original distribution P_0. Another k samples are generated by humans interacting with the newest model f_{t-1} (e.g. users auto-tag new images). The dataset, derived from accumulated online content, thus evolves as

$$\mathcal{S}_t = \mathcal{S}_{t-1} \cup \{(x_i, y_i)\}_{i \in [m]} \cup \{(x_j, f_{t-1}(x_j)\}_{j \in [k]},$$

with $(x_i, y_i) \overset{\text{iid}}{\sim} \mathsf{P}_0$ and $x_j \overset{\text{iid}}{\sim} \mathsf{P}_0(x)$, where $\mathsf{P}_0(x)$ denotes the marginal over the covariates. The model is then updated by re-training on the growing dataset, $f_t \sim \mathcal{A}(\mathcal{S}_t)$. Formally, the data feedback model we instantiate in our experiments is defined in Algorithm 1.

Our overall goal is to analyze the behavior of f_t over time. Concretely, we are concerned with *bias amplification*, tracked via a particular bias statistic $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. We will measure the expected difference between the bias of the initial, human-annotated distribution P_0 and the bias of the model f_t. Thus, in both our theoretical and empirical analyses, we will measure amplification as

$$\left| \mathbb{E}_{f_t}\left[\mathbb{E}_{(x,y)\sim P_0}\left[\phi(x, y) - \phi(x, f_t(x))\right]\right] \right|$$

over time t. The expectation in this bias term, $\mathbb{E}_{f_t}[\cdot]$, is an expectation over all random objects up to time t, which includes random draws in each dataset $\mathcal{S}_t$ and random draws of the model f_t.

One important aspect of this setting is that all covariates are sampled from the same distribution $\mathsf{P}_0(x)$, which remains fixed over time. This assumption is natural in situations similar to Figure 1, where

predictions of the image-tagging model may not influence the types of photos taken. Though we make this choice to simplify our analysis, this setting still poses challenging tradeoffs; in Section 5.1, we show that retraining classifiers with future data improves accuracy at the cost of increasing bias.

4 STABILIZING BIAS AMPLIFICATION

In this section, we develop theoretical tools that will allow us to make predictions about bias amplification in experimental settings. We begin with a toy example connecting sampling to calibration in Section 4.1, show how calibration leads to stability via bounds on amplification in Section 4.2, and finally discuss prior work showing when calibration naturally arises for models in Section 4.3.

4.1 ILLUSTRATIVE EXAMPLE

We begin with an example to emphasize how data feedback may become unstable. For clarity, an illustration of this example is provided in Figure 5 (in Appendix B). Consider a set of images of female healthcare workers with high inherent uncertainty – they could each be either a doctor or a nurse, depending on context cues that are not present in the image (Figure 5 left). In this case, data feedback on a dataset with twice as many nurses as doctors can rapidly destabilize.

More concretely, any Bayes optimal classifier would predict new examples only as nurse, as nurses are the majority class and the image is indistinguishable otherwise. Such a model would exacerbate the nurse bias in the dataset (Figure 5 top). A natural solution solution to this problem would be to predict nurses and doctors at a rate equal to the original distribution. Specifically, a sampling-based model trained to reproduce the training distribution would exhibit this behavior, continuing to label a random $\frac{2}{3}$ of the examples as nurses. Though such a model may have less utility, it would maintain the level of nurse bias in the dataset (Figure 5 bottom).

A training algorithm that produces models whose outputs match the bias of the training distribution is said to be consistently calibrated, and we will now formally define and connect calibration to stability.

4.2 ACHIEVING STABILITY THROUGH CALIBRATION

Setup and goals. We first define a few objects useful for analysis. We call the number of training samples at time t as $n_t := n_{t-1} + m + k = n_0 + t(m+k)$. A mixture of past training data, new human-annotated data, and new model-annotated data, the training data distribution at time t is

$$\mathsf{P}_t = \frac{n_{t-1}}{n_t}\mathsf{P}_{t-1} + \frac{m}{n_t}\mathsf{P}_0 + \frac{k}{n_t}\widehat{\mathsf{P}}_0(f_{t-1}),$$

where the shorthand $\widehat{\mathsf{P}}(f)$ denotes the model-annotated distribution, which is the *relabeling* of P by f. Samples are drawn from this distribution by first sampling a covariate $x \sim \mathsf{P}(x)$ and then returning the annotated pair $(x, f(x))$. Additionally, for ease of analysis in this section, we study the case where the dataset $\mathcal{S}_t$ is drawn fresh from its corresponding distribution P_t at every time, $\mathcal{S}_t \sim \mathsf{P}_t^{n_t}$ [1].

As a reminder, our interest is in the expected bias amplification of a learning algorithm $\mathcal{A}$ at time t,

$$\left|\mathsf{P}_0\phi - \mathbb{E}_{f_t}\left[\widehat{\mathsf{P}}_0(f_t)\phi\right]\right| := \left|\mathbb{E}_{f_t}\left[\mathbb{E}_{(x,y)\sim\mathsf{P}_0}\left[\phi(x,y) - \phi(x, f_t(x))\right]\right]\right|,$$

where the left hand side uses the shorthand $\mathsf{P}\phi := \mathbb{E}_{(x,y)\sim\mathsf{P}}[\phi(x,y)]$, defined as expectation of the bias metric ϕ over distribution P.

Calibration. In the previous nurses versus doctors example, we discovered that a model that faithfully represented the training data distribution was more stable under data feedback. Now, we formalize what it means to faithfully represent the data distribution: We say a learning algorithm is *consistently calibrated* if the bias of the model is similar to the bias of the training distribution.

Definition 1 (Consistent Calibration)**.** *A learning algorithm $\mathcal{A}$: $(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}$ is (δ, ϕ, $\mathsf{P}(x)$, n)-consistently calibrated if, for any joint distribution $\mathsf{Q}(x, y)$ with marginal $\mathsf{P}(x)$,*

$$\left|\mathbb{E}_{\mathcal{S}\sim\mathsf{Q}^n, f\sim\mathcal{A}(\mathcal{S})}\left[\mathsf{Q}\phi - \widehat{\mathsf{Q}}(f)\phi\right]\right| \leq \delta.$$

[1] This generative model assumes $\mathcal{S}_t$ is a new draw from P_t at each timestep, which differs from the definition in Algorithm 1 where $\mathcal{S}_t$ is constructed by concatenating new samples with the prior timestep's dataset. We make this simplifying assumption only for the theoretical analysis in this section since we are interested in the dependence between deployed models and training data distributions, not in the dependence introduced by the draw of each dataset. We expect this difference in definition to be small as the sample size grows large.

If a learning algorithm is consistently calibrated, it means that in expectation, the bias of the trained model will be close to the dataset bias. Since this condition holds for all joint distributions sharing a marginal, and since the covariate marginal does not change does not change during data feedback (i.e. $\mathsf{P}_t(x) = \mathsf{P}_0(x)$ for all t), if learning algorithm $\mathcal{A}$ is consistently calibrated for the initial distribution $\mathsf{P}_0(x)$, $\mathcal{A}$ will also be consistently calibrated for all $\mathsf{P}_t(x)$ (formalized in Lemma C.1 in Appendix C). In fact, this property naturally arises in some settings, as discussed in the next subsection.

Intuitively, satisfying this definition helps to control bias amplification. At time t, a consistently calibrated algorithm $\mathcal{A}$ will have bias no more than δ greater than its training distribution P_t. In turn, the bias of P_t is reduced when adding human-annotated samples and increased when adding model-annotated samples. The balance of these two quantities is crucial for stabilizing bias amplification.

Stability. Our main feedback stability result is a direct consequence of consistent calibration.

Theorem 1 (Feedback Stability)**.** *Let $\mathcal{A}$: $(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}$ be a (δ_n, ϕ, $\mathsf{P}_0(x)$, n)-consistently calibrated learning algorithm, where calibration error δ_n is a monotone non-increasing function of dataset size n. Then, under the data feedback procedure, for all time t,*

$$\left|\mathbb{E}_{f_t}\left[\mathsf{P}_0\phi - \widehat{\mathsf{P}}_0(f_t)\phi\right]\right| \leq \left(1 + \sum_{i=1}^{t} \frac{k}{n_i} \prod_{j=i+1}^{t} \frac{n_j - m}{n_j}\right) \delta_{n_0} \leq \frac{m+k}{m}\delta_{n_0}.$$

The proof is provided in Appendix C. Surprisingly, the bound shows that data-driven feedback can be stable even in the limit of $t \to \infty$. From inspecting the simplified upper bound, it is clear that both a larger number of human-annotated examples m and a smaller initial calibration error δ_{n_0} serve to stabilize the system and minimize bias amplification. This leads to a natural question: in which situations can we expect a small consistent calibration error?

Intuitively, models that behave like samplers will have low calibration error. In particular, suppose that model f_t has accurately learned the conditional distribution of P_t, i.e. $d_{TV}(\mathsf{P}_t(y|x), f_t(y|x)) \leq \delta$. Now, we perform a comparison of two prediction strategies commonly used in machine learning: sampling $y \sim f_t(y|x)$ and argmax prediction $y = \operatorname{argmax}_y f_t(y|x)$.

If labels are sampled, $y \sim f_t(y|x)$, then $d_{TV}(\mathsf{P}_t, \widehat{\mathsf{P}}(f_t)) \leq \delta$ by definition, and so f_t is δ-calibrated for any metric ϕ by post-processing. Alternatively, if the top prediction $y = \operatorname{argmax}_y f_t(y|x)$ is used, f_t is not guaranteed to be δ-calibrated for bias metric ϕ, similar to the example in Figure 5.

While it is unsurprising that sampling maintains calibration and argmax predictions can be miscalibrated, prior work has discovered that certain models which do not explicitly sample can still behave like samplers (Nakkiran & Bansal, 2020) and thus also provide feedback stability.

4.3 Achieving calibration through Distributional Generalization

As in the example in Figure 5, when there is large uncertainty over the true labels (doctors versus nurses), one strategy for reducing bias is to sample according to the training distribution. Distributional Generalization (DG) (Nakkiran & Bansal, 2020) demonstrates that interpolating classifiers, which are argmax predictors, behave similarly; when the model has high uncertainty over the true labels, it produces outputs that mimic the training distribution.

Concretely, let $L : \mathcal{X} \to [m]$ be a partioning of the input space into $m \in \mathbb{Z}_+$ parts, where similar points with high uncertainty are grouped together. This partitioning "coarsens" the input space by mapping hard-to-learn regions to single points. DG finds that at this level of coarseness, samples labeled by interpolating classifiers look like samples from the training distribution, i.e. $(L(x), f(x)) \approx (L(x), y)$ (Nakkiran & Bansal, 2020). That is, *within a specific partition*, the random process of drawing a sample x and labeling it with a deterministic classifier $y = f(x)$ produces a distribution similar to drawing x and then sampling a label from the true conditional $y \sim p(y|x)$.

If the bias metric ϕ was applied over this coarsened space, we may expect feedback stability as a natural consequence of model outputs behaving like samples. We will now formalize this intuition by linking DG to consistent calibration. To do so, we first define an appropriate input partioning L.

Definition 2 (Distinguishable Feature (Nakkiran & Bansal, 2020))**.** *Let $L : \mathcal{X} \to [m]$ be a coarsening of the input domain $\mathcal{X}$ into $m \in \mathbb{Z}_+$ parts. Define $\widehat{\mathsf{P}}(L)$ as the relabeling of P by L. Then, L is a (δ, $\mathcal{A}$, $\mathsf{P}(x)$, n)-distinguishable feature if*

$$\mathbb{P}_{\mathcal{S}\sim\widehat{\mathsf{P}}(L)^n, f\sim\mathcal{A}(\mathcal{S}), x\sim\mathsf{P}(x)}\left[f(x) = L(x)\right] \geq 1 - \delta.$$

The partitioning L defines how points in P are grouped together. An appropriate partioning is one where the learner $\mathcal{A}$ can classify the group identity of each point with high accuracy. Additionally, note that the coarsening L does not depend on the label distribution and relies only on the marginal $\mathsf{P}(x)$. This property is important for data feedback; if L is distinguishable for the initial distribution P_0, it will continue to be distinguishable for all P_t.

With an appropriate partitioning defined, we can now connect it to consistent calibration via DG.

Lemma 4.1. *Suppose that bias metric ϕ is a function of a (δ, $\mathcal{A}$, $\mathsf{P}(x)$, n)-distinguishable feature L, i.e. $\phi(x, y) = T(L(x), y)$ for some bounded $T : [m] \times \mathcal{Y} \to \mathbb{R}$. Then, under DG (Conjecture 1 in Appendix C.3), learning algorithm $\mathcal{A}$ is (δ, ϕ, $\mathsf{P}(x)$, n)-consistently calibrated.*

The proof is provided in Appendix C.4. This lemma is an immediate consequence of DG (Conjecture 1), which states that the coarsened model outputs $(L(x), f(x))$ are similar to the coarsened training data $(L(x), y)$ for all bounded tests T; this is the basis for the statement that model outputs behave like samples, i.e. $(L(x), f(x)) \approx (L(x), y)$. The given bias metric ϕ is simply one such test.

This result shows that under DG, global stability can be achieved (excess bias bounded by $\frac{m+k}{m}\delta_{n_0}$ for all time) if the bias metric ϕ is a function of a δ_{n_0}-distinguishable feature on the initial dataset.

4.4 Instantiating feedback upper bounds in experiments

We have seen two strategies for consistent calibration: 1) explicitly, through estimating the conditional distribution well and sampling outputs, and 2) implicitly through DG, where interpolating classifiers provide guarantees as long as the bias metric is a function of a sufficiently coarse input feature.

In these settings, one more condition is needed for Theorem 1 to apply – that calibration errors δ_n are non-increasing with dataset size n. Although not guaranteed, many learning algorithms and natural data distributions satisfy this property experimentally, especially if regularization is tuned (Nakkiran et al., 2020), as in done in practice. We therefore believe it is a reasonable assumption to expect calibration error to be a monotone decreasing function of dataset size in most experimental situations.

In the next section, we will explore how our derived predictions can help estimate bias amplification in realistic data feedback settings. In order to instantiate the bound in Theorem 1, we need to know the initial consistent calibration error δ_{n_0}. As a practical approximation, we estimate δ_{n_0} empirically via the calibration error of the initial model f_0. Although this empirical estimate is a lower bound on the consistent calibration error, we find that it is a useful guide, and we observe that the corresponding predictions from Theorem 1 still bound the empirical amplification.

5 Tracking bias amplification in feedback experiments

We consider three natural real-world settings that give rise to data feedback: image classification, visual role-labeling, and conditional language generation. The image classification and visual role-labeling settings are inspired by the example in Figure 1, where existing biases in image annotations may amplify. The language modeling setting is inspired by the rise of online conversational agents (Dinan et al., 2021) and assisted story writing systems (Donahue et al., 2020), for which there are real concerns about model-generated toxicity or bias (Sheng et al., 2019).

In each of these cases, we will study the behavior of data feedback in three steps: instantiate Algorithm 1, measure the empirical bias amplification, and then compare with the predictions of Theorem 1. Our experiments identify that feedback stability arises when models behave like samplers and calibration error is small. For each setting, we describe the main experimental setup followed by the results. Extra setup details are in Appendix F, and corresponding ablations are in Appendix G.

5.1 Image classification

Setting up the label bias experiment.

Studying data feedback over many rounds requires very large datasets, and we use the CIFAR-5m dataset (Nakkiran et al., 2021), which contains 5 million synthetically generated examples. We re-balance the dataset to contain 50% dogs, resulting in a 9:1 imbalance ratio compared to any other class. For our bias metric ϕ, we track the fraction of the model's predictions that are dogs. Ideally, we would like this fraction to remain near 50%, the true data distribution level.

For the model, we train a BaiduNet9 (Li et al., 2019) on the growing dataset from scratch at each timestep, and hyperparameters are re-tuned every time. We run data feedback (Algorithm 1) with an

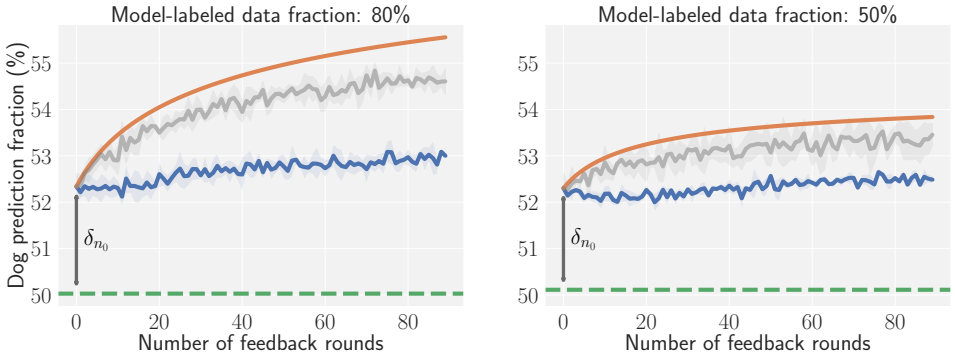

Figure 2: Results of data feedback (Algorithm 1) on CIFAR with dog imbalance. Bias is measured as the fraction of model predictions that are dogs. Empirical trends are shown with the mean and standard deviation over 3 random seeds. **Blue:** Empirical trend, BaiduNet9 trained from scratch at each round. **Orange:** Amplification upper bound (Theorem 1) for the blue trend, with δ_{n_0} estimated empirically. **Gray:** Worst-case empirical setting (details in Appendix D.1). **Takeaways**: The empirical curves qualitatively match the bounds, with bias amplifying more with more model-labeled samples. In both cases, the orange upper bounds the empirical trends.

initial dataset size $n_0 = 50$k and new samples per round $m + k = 5$k. We report results both when 80% and 50% of new samples are model-labeled each round ($\frac{m+k}{m} = 5$ and 2 respectively).

Analyzing label bias amplification.

We show the results of running data feedback on the CIFAR-5m dataset in Figure 2 (blue trend). As predicted by Theorem 1, the fraction of model predictions which are dogs grows faster in the setting with a greater fraction of model-labeled samples. Specifically, the bias amplifies $+0.8\%$ when $\frac{m+k}{m} = 5$ (left) and $+0.3\%$ when $\frac{m+k}{m} = 2$ (right). We observe that the theoretical bounds, though conservative, are consistent with the empirical results. This matches our expectations, since prior work suggests that Distributional Generalization holds for CIFAR classifiers and that the dog class is a distinguishable feature (Nakkiran & Bansal, 2020), which by Lemma 4.1 implies stability.

While in both settings the dog bias amplifies, the overall classification accuracies of the models improve throughout data feedback, a result of increasing dataset size. Specifically, as the size of the training set grows from $n_0 = 50$k to $n_{90} = 500$k over 90 rounds of data feedback, average classification accuracy improves $+2.4\%$ and $+1.6\%$ for the models with 50% and 80% model-labeled samples (Figure 6 in Appendix E.1). Trading off this increase in utility with greater label bias is a challenge for model developers who seek to update their models to new data. Our theoretical bounds take a step towards characterizing this tradeoff by upper bounding empirical bias amplification.

Lastly, we discuss the source of the looseness in our bounds and present a more rigorous test of our upper bound with a worst-case setting in Appendix D.1. The results are displayed in the gray trend in Figure 2; we note that our bounds qualitatively capture the empirical behavior in this setting well.

5.2 Visual role-labeling

Setting up the gender bias experiment.

We run data feedback on the imSitu dataset (Yatskar et al., 2016), where models are asked to predict both the verb category of an image (e.g. cooking, jumping, etc.) as well as labels for the subjects and objects (e.g. female, basketball, etc.). Zhao et al. (2017) found that models trained on this dataset amplify gender disparities at test-time; for example, 67% of cooking category images in the dataset are labeled female, but a ResNet18 trained on the dataset will label 84% of cooking images as female. Based on this observation, we select the verb categories with an existing female gender bias, and we measure the fraction of the model's predictions that are labeled female over these verbs [2].

We train the default ResNet18 (He et al., 2016) conditional random fields model from scratch at each timestep, and hyperparameters are re-tuned every time. We run data feedback (Algorithm 1) with an initial dataset size $n_0 = 50$k and new samples per round $m + k = 5$k. We report results both when 80% and 50% of new samples are model-labeled each round ($\frac{m+k}{m} = 5$ and 2 respectively).

[2] Specifically, in Figure 3, we consider the verb categories where the dataset female label ratios lie between 60% to 80%. In Appendix G.2, we provide plots for 0-20%, 20-40%, 40-60%, and 80-100%.

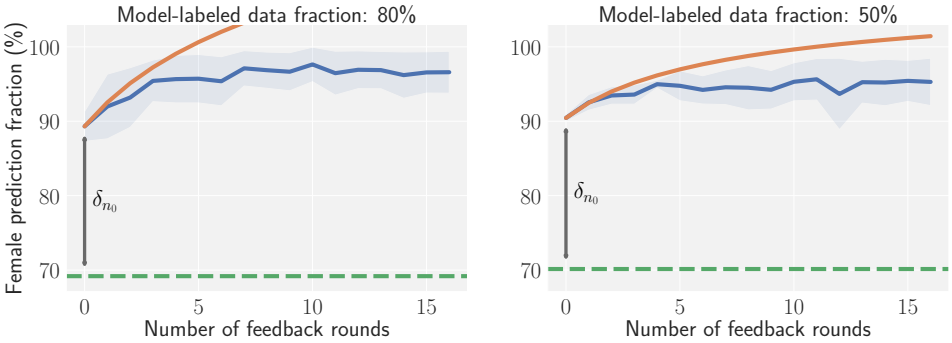

Figure 3: Results of data feedback (Algorithm 1) on the imSitu dataset. Bias is measured as the fraction of predictions that are labeled as female within the verb categories that have an existing female bias. **Blue:** Empirical trend, ResNet18 trained from scratch at each round, shown with the mean and standard deviation over 3 random seeds. **Orange:** Amplification upper bound (Theorem 1), with δ_{n_0} estimated empirically. **Takeaways**: Since the initial calibration error δ_{n_0} is large, the bounds quickly become vacuous (crossing over the 100% female prediction fraction mark), which is mirrored by the empirical bias also reaching near 100%.

Analyzing gender bias amplification.

We show results of data feedback on the imSitu dataset in Figure 3. The initial calibration error δ_{n_0} is much larger than in the CIFAR setting; the initial trained model predicts females 90% of the time, though the dataset female fraction level is at 70%. As a result, the bound from Theorem 1 quickly becomes vacuous, crossing over the 100% female prediction fraction mark. This prediction is mirrored by the empirical bias also reaching near 100% in just 16 rounds of feedback (97% and 95% female prediction fraction when 80% and 50% of new samples are model-labeled, respectively).

Male prediction bias is also amplified on this task. In Figure 7 in Appendix E.2, we plot the male prediction bias over the verb categories with an existing male skew for these same models and find that it amplifies quickly, similar to Figure 3. Interestingly, this implies that gender biases quickly amplify simultaneously and in both directions; for female-biased categories, predictions become more female, and for male-biased categories, predictions become more male.

5.3 Conditional Language Modeling

Setting up the toxicity and repetition bias experiment.

We use the Real Toxicity Prompts dataset (Gehman et al., 2020), which is a set of 100k sentences collected from the Open-WebText Corpus (Gokaslan & Cohen, 2019) with varying levels of toxicity. Each sentence was split into two halves, a prompt and a continuation. We use this to construct a language modeling task where a model is asked to complete a sentence given a prompt.

We measure two bias metrics on the model output: toxicity and repetition. Toxicity is measured by the fraction of model outputs classified as toxic by the Detoxify classifier (Hanu & Unitary team, 2020). We also measure a specific form of repetition bias: the average number of quotation marks in the generated text. Repetitive text is a common degeneracy of language models (Holtzman et al., 2020; Fan et al., 2018), and we count quote frequencies as a simple approximation after observing that repetitive outputs in this setting commonly contained many quotes (see Appendix E.3 for examples).

We finetune a pretrained GPT-2 small (Radford et al., 2019) at each round, with hyperparameters retuned every time. To generate new sentence completions, we consider two common schemes: nucleus sampling (Holtzman et al., 2020) (`top_p` $= 0.9$) and beam search (Graves, 2012) (`num_beams` $= 10$). We run data feedback (Algorithm 1) with an initial dataset size $n_0 = 20$k and new samples per round $m + k = 5$k, with 80% being model-labeled ($k = 4$k).

Analyzing toxicity and repetition bias amplification.

Figure 4 shows the results of data feedback on the Real Toxicity Prompts dataset. Comparing beam search (blue) to nucleus sampling (brown), the toxicity of the final nucleus sampling models (14.5%) did not change from their initial level. However, the toxicity of the final beam search models (11.5%) decreased by about 3% from their initial level; in this case, beam search amplified the toxicity bias downward since the initial model's toxicity (14.5%) was lower than the dataset toxicity level (23%).

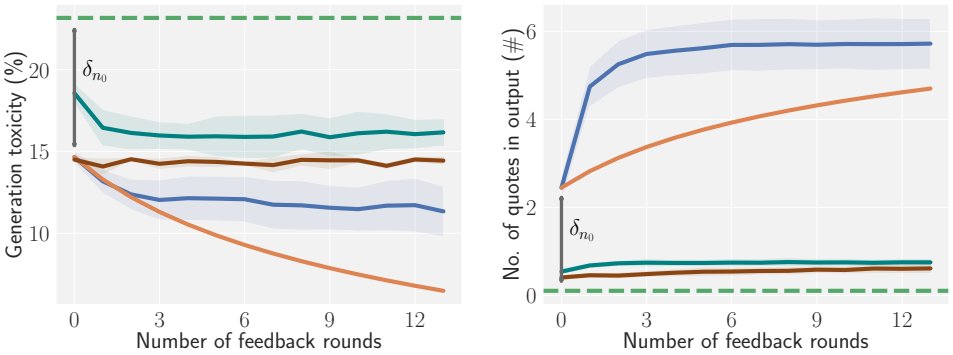

Figure 4: Results of data feedback (Algorithm 1) on the Real Toxicity Prompts dataset (Gehman et al., 2020). Empirical trends are shown with the mean and standard deviation over 3 random seeds. Bias is measured in two ways; **left:** the fraction of model outputs that are classified as toxic by a separate toxicity classifier (toxicity bias), and **right:** the average number of quotation marks in the generated text (repetition bias). **Blue:** Finetuned GPT2-small with beam search outputs. **Orange:** Amplification upper bound (Theorem 1) for the blue trend, with δ_{n_0} estimated empirically. **Brown:** Finetuned GPT2-small with nucleus sampling outputs. **Teal:** Proposed intervention of finetuned and overfit GPT2-small with beam search outputs. **Takeaways**: Nucleus sampling is more stable than beam search for both bias metrics, particularly for repetition bias, demonstrating that sampling is more stable than argmax predictions. The proposed intervention of overfit beam search largely resolves the issues with beam search; the empirical curves behave much more similarly to nucleus sampling with respect to both toxicity bias and especially repetition bias, demonstrating the stabilizing effect of the intervention.

Repetition bias results paint a more dramatic difference between the two. While the average number of quotes in generated text increases little for nucleus sampling (0.4 to 0.6), it amplifies significantly for beam search (2.5 to 5.7). In fact, the beam search empirical amplification even exceeds Theorem 1's upper bound. We believe this is due to the lack of a calibration guarantee, since Distributional Generalization has not been shown to hold for language models (and thus Lemma 4.1 cannot guarantee stability). In its absence, the argmax-style generation strategy of beam search is exacerbating the existing repetition bias, in line with the sampling vs argmax stability analysis in Section 4.

An intervention to stabilize toxicity and repetition bias.

We now test our understanding of bias amplification by designing an intervention to mitigate amplification for beam search models. Leveraging the claim in Distributional Generalization that interpolating models behave like samplers, we overfit the beam search model with the goal of making it interpolate the training data. The intervention is simple: we finetune the GPT2-small model for 5 times the number of training epochs as before. This dropped the round 0 training loss from 3.5 to 0.4, and the test set perplexity correspondingly jumped from 32 to 599.

Figure 4 (teal) shows the results of the intervention. Overfitting significantly improves the stability of the beam search model; the average number of quotes output by the final model is reduced from 5.7 to 0.8, which is closer to the nucleus sampling level at 0.6. The relative amplification was also reduced, as the final overfit beam search model was only $1.4\times$ as repetitive as the initial model, down from a $2.3\times$ relative amplification before. Sample outputs of all three models are in Appendix E.3.

In Appendix D.2, we discuss the utility of this intervention, investigating whether the overfit model is simply memorizing the training data by counting n-gram frequencies. Regardless, our experimental results are consistent with our earlier theoretical characterizations of stability and suggest that approaches for improving calibration may be broadly useful for mitigating bias amplification.

6 Conclusion

We propose a new setting called *data feedback*, where past model outputs act as training data in the future. We show that the natural decision to retrain a deployed model can increase utility while also amplifying biases. We then provide conditions for stability (namely, consistent calibration) and derive corresponding upper bounds on bias amplification. The utility of these predictions is realized by experiments in image classification, visual role-labeling, and language modeling, which confirm the observation that sampling-like behaviors often result in better calibration and greater feedback stability. Finally, we leverage our insight to design a mitigation strategy for unstable feedback systems. We hope our work will encourage further discussion around mitigation and prevention strategies.

REFERENCES

G. A. Adam, C.-H. K. Chang, B. Haibe-Kains, and A. Goldenberg. Hidden risks of machine learning applied to healthcare: unintended feedback loops between models and future data causing model degradation. In *Machine Learning for Healthcare Conference*, 2020.

G. A. Adam, C.-H. K. Chang, B. Haibe-Kains, and A. Goldenberg. Error amplification when updating deployed machine learning models. In *Machine Learning for Healthcare Conference*, 2022.

L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, and A. Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

G. Brown, S. Hod, and I. Kalemaj. Performative prediction in a stateful world. In *Artificial Intelligence and Statistics (AISTATS)*, 2022.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

A. J. Chaney, B. M. Stewart, and B. E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *ACM Conference on Recommender Systems*, 2018.

S. Coate and G. C. Loury. Will affirmative-action policies eliminate negative stereotypes? In *The American Economic Review*, 1993.

P. Dandekar, A. Goel, and D. T. Lee. Biased assimilation, homophily, and the dynamics of polarization. In *Proceedings of the National Academy of Sciences*, 2013.

L. N. Darlow, E. J. Crowley, A. Antoniou, and A. J. Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

J. Dhamala, T. Sun, V. Kumar, S. Krishna, Y. Pruksachatkun, K. Chang, and R. Gupta. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 862–872, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097.

E. Dinan, A. Fan, A. Williams, J. Urbanek, D. Kiela, and J. Weston. Queens are powerful too: Mitigating gender bias in dialogue generation. *arXiv preprint arXiv:1911.03842*, 2019.

E. Dinan, G. Abercrombie, A. S. Bergman, S. Spruit, D. Hovy, Y. Boureau, and V. Rieser. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*, 2021.

C. Donahue, M. Lee, and P. Liang. Enabling language models to fill in the blanks. In *Association for Computational Linguistics (ACL)*, 2020.

D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian. Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847*, 2017.

A. Fan, M. Lewis, and Y. Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia. A brief review of domain adaptation. In *Advances in Data Science and Information Engineering*, 2021.

L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Empirical Methods in Natural Language Processing*, 2020.

A. Gokaslan and V. Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Y. Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2004.

A. Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

M. Hall, L. van der Maaten, L. Gustafson, and A. Adcock. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022.

L. Hanu and Unitary team. Detoxify. https://github.com/unitaryai/detoxify, 2020.

M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters. Strategic classification. In *ACM Conference on Innovations in Theoretical Computer Science*, 2016.

T. B. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*, 2018.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2020.

Z. Izzo, L. Ying, and J. Zou. How to learn when data reacts to your model: performative gradient descent. In *International Conference on Machine Learning (ICML)*, 2021.

R. Jiang, S. Chiappa, T. Lattimore, A. György, and P. Kohli. Degenerate feedback loops in recommender systems. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2019.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

A. Kumar, T. Ma, and P. Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning (ICML)*, 2020.

D. Kumar, P. G. Kelley, S. Consolvo, J. Mason, E. Bursztein, Z. Durumeric, K. Thomas, and M. Bailey. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS)*, 2021.

K. Leino, E. Black, M. Fredrikson, S. Sen, and A. Datta. Feature-wise bias amplification. In *International Conference on Learning Representations (ICLR)*, 2019.

B. Li, Z. Cheng, and Y. Bao. Cifar10 training baidunet9. https://github.com/BAIDU-USA-GAIT-LEOPARD/CIFAR10-Training-BaiduNet9, 2019.

Z. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning (ICML)*, 2018.

L. T. Liu, A. Wilson, N. Haghtalab, A. T. Kalai, C. Borgs, and J. Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Conference on Fairness, Accountability, and Transparency*, 2020.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

J. P. Miller, J. C. Perdomo, and T. Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning (ICML)*, 2021.

P. Nakkiran and Y. Bansal. Distributional generalization: A new kind of generalization. *arXiv preprint arXiv:2009.08092*, 2020.

P. Nakkiran, P. Venkat, S. Kakade, and T. Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

P. Nakkiran, B. Neyshabur, and H. Sedghi. The deep bootstrap framework: Good online learners are good offline generalizers. In *International Conference on Learning Representations (ICLR)*, 2021.

Y. Ouali, C. Hudelot, and M. Tami. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020.

J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. In *International Conference on Machine Learning (ICML)*, 2020.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. https://cdn.openai.com/papers/whisper.pdf, 2022.

J. S. Rosenfeld. *Scaling laws for deep learning*. PhD thesis, Massachusetts Institute of Technology, 2021.

M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. A. Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Association for Computational Linguistics (ACL)*, 2022.

S. Schmit and C. Riquelme. Human interaction with recommendation systems. In *Artificial Intelligence and Statistics (AISTATS)*, 2018.

E. Sheng, K. Chang, P. Natarajan, and N. Peng. The woman worked as a babysitter: On biases in language generation. In *Empirical Methods in Natural Language Processing*, 2019.

R. Shu, H. H. Bui, H. Narui, and S. Ermon. A dirt-t approach to unsupervised domain adaptation. In *International Conference on Learning Representations (ICLR)*, 2018.

A. Sinha, D. F. Gleich, and K. Ramani. Deconvolving feedback loops in recommender systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

H. Steck. Calibrated recommendations. In *ACM Conference on Recommender Systems*, 2018.

Jigsaw team. Toxic comment classification challenge. https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/overview, 2018.

Jigsaw team. Jigsaw unintended bias in toxicity classification. https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification, 2019.

Jigsaw team. Jigsaw multilingual toxic comment classification. https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification, 2020.

B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. Yfcc100m: The new data in multimedia research. In *Communications of the Association for Computing Machinery (ACM)*, 2016.

A. Venugopal, J. Uszkoreit, D. Talbot, F. J. Och, and J. Ganitkevitch. Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Empirical Methods in Natural Language Processing*, 2011.

T. Wang, J. Zhao, M. Yatskar, K. Chang, and V. Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *International Conference on Computer Vision (ICCV)*, 2019.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

M. Yatskar, L. Zettlemoyer, and A. Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

J. Zhao, T. Wang, M. Yatskar, V. Ordóñez, and K. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

Z. Zhu, T. Luo, and Y. Liu. The rich get richer: Disparate impact of semi-supervised learning. *arXiv preprint arXiv:2110.06282*, 2021.