# Bias in Intent Detection: A Dynamical Systems Perspective

EDUARDO SANCHEZ-KARHUNEN, University of Seville, Spain
JOSE F. QUESADA-MORENO, University of Seville, Spain
MIGUEL A. GUTIÉRREZ-NARANJO, University of Seville, Spain

Intent detection is a critical task in natural language processing (NLP), powering applications such as chatbots and dialogue systems. Although deep learning models have greatly improved intent classification, their internal mechanisms remain poorly understood, raising concerns about transparency and fairness. Recent studies have applied dynamical systems theory to analyze RNNs by examining their internal state dynamics. We propose a novel bias evaluation framework that examines sentence trajectories within the model's state space. By analyzing the dynamic evolution of hidden states and their final alignment with decision-making layers, we identify key mechanisms for defining new bias metrics: trajectory sparsity, final state clustering, and readout vector alignment. This interpretable framework offers a principled approach to diagnosing and mitigating bias.

Keywords: Natural Language Processing, Intent Detection, Dynamical Systems, Bias, Interpretability, Transparency, Fairness

**Reference Format:**
Eduardo Sanchez-Karhunen, Jose F. Quesada-Moreno, and Miguel A. Gutiérrez-Naranjo. 2025. Bias in Intent Detection: A Dynamical Systems Perspective. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 7 pages.

## 1 Introduction

Intent detection is a foundational task in NLP, serving as a core operation in applications such as virtual assistants and dialogue systems [24]. The tasks involve identifying and classifying the purpose or operational goal (i.e. intent) behind an user's query, mapping natural language expressions to predefined categories. It plays a crucial role in high-stakes domains including healthcare, education, and access to essential services, where incorrect interpretation of user intent can lead to serious consequences.

Deep learning models, particularly modern RNNs, are widely used and have significantly improved performance in various NLP tasks, such as sentiment analysis [19] and intent detection [13]. Despite their success, RNNs inner workings remain largely opaque, posing challenges for both interpretability and fairness. This lack of transparency arises from their nonlinear nature and the high dimensionality of their hidden layers.

A new reverse engineering paradigm has emerged to analyze RNNs through the lens of dynamical systems theory [28]. Rather than focusing on individual neurons or gates, this approach treats the network as a dynamical system where sentences evolve as trajectories within a high-dimensional state space. This perspective has provided

Authors' Contact Information: Eduardo Sanchez-Karhunen, University of Seville, Seville, Spain, fesanchez@us.es; Jose F. Quesada-Moreno, University of Seville, Seville, Spain, jquesada@us.es; Miguel A. Gutiérrez-Naranjo, University of Seville, Seville, Spain, magutier@us.es.


*EWAF'25, June 30–July 02, 2025, Eindhoven, NL*

valuable insights into the computational mechanisms underlying various NLP tasks, including text classification [1], sentiment analysis [20], and intent detection [26].

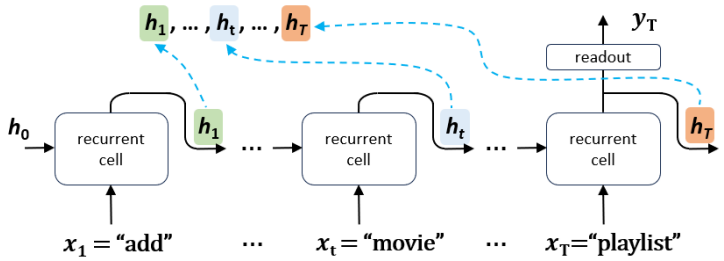
In this work, we extend this approach to the study of fairness and bias in NLP [3], with a particular focus on intent detection. To the best of our knowledge, this is the first study to apply dynamical system theory to audit and mitigate bias in NLP models. We propose three fairness metrics based on state space dynamics to assess disparities in sentences trajectories across different sociodemographic and linguistic groups. Our framework provides an interpretable method for diagnosing and mitigating bias in intent detection models.

## 2 RNNs: A Dynamical Systems Approach

Computations performed by RNNs can be described using Eqs. (1) and (2), where $t$ typically represents time, $\mathbf{h}_t \in \mathbb{R}^n$ denotes the *hidden state*, and $\mathbf{x}_t \in \mathbb{R}^m$ is the external input at step $t$. The nonlinear update function $\mathbf{F}$ governs the evolution of the hidden state, depending on the specific recurrent cell architecture used [27].

As shown in Fig. 1, given an input sequence $\mathbf{x}_1, \ldots, \mathbf{x}_T$, the network updates its hidden state $\mathbf{h}_t$ at each step $t$ based on the previous hidden state $\mathbf{h}_{t-1}$ and the current input $\mathbf{x}_t$. Predictions $\mathbf{y}_t$ are obtained by passing the hidden states through a linear *readout layer*, where $\mathbf{W}$ is a $n \times n$ readout weight matrix and $\mathbf{b}$ is a bias vector. Each row $\mathbf{r}_i$ of $\mathbf{W}$ is called a readout vector. In classification tasks, the output $\mathbf{y}_t$ consists of $N$ logits, one for each class label. In many-to-one contexts (e.g. intent detection or sentiment analysis), only the final hidden state $\mathbf{h}_T$ is used to generate a single prediction $\mathbf{y}_T$.

Systems governed by difference equations as in Eq. 1 are called *discrete-time dynamical systems*, with their state represented by $\mathbf{h}_t$ at time $t$. In the context of RNNs, the update function $\mathbf{F}$ is inherently nonlinear (e.g. GRU [7], LSTM [15], Vanilla [11]), which classifies these systems as nonlinear discrete-time dynamical systems [16].




$$\mathbf{h}_t = \mathbf{F}(\mathbf{h}_{t-1}, \mathbf{x}_t) \quad (1)$$

$$\mathbf{y}_t = \mathbf{W}\mathbf{h}_t + \mathbf{b} \quad (2)$$

Fig. 1. Sequence of hidden states $\mathbf{h}_1, \ldots, \mathbf{h}_T$ associated with an input sentence $\mathbf{x}_1, \ldots, \mathbf{x}_T$ as it is processed by an RNN. The hidden states $\mathbf{h}_t$ evolve in response to input tokens $\mathbf{x}_t$, representing the progression of the network's internal dynamics. The final hidden state $\mathbf{h}_T$ captures the information of the input sequence and is used to generate the prediction $\mathbf{y}_T$ via the readout layer.

## 3 Internal Dynamics of Intent Detection

Traditional dynamical system analysis treats individual neurons as system parameters, but the high dimensionality of RNNs presents significant challenges for standard state-space analysis [14]. To address this, recent studies have adopted a higher-level perspective, reverse-engineering the computational mechanisms learned by RNNs [28]. These approaches have been successfully applied to tasks such as sentiment analysis [20] and text classification [1], revealing interpretable geometric and dynamical structures within the hidden state space.

This line of research was extended to the problem of intent detection, a common task in NLP. For this purpose, the SNIPS dataset [8] was employed. It consists of short, natural language sentences such as "Play some jazz music" or "Book a table for two," each annotated with a specific intent class like *PlayMusic*, *BookRestaurant*, or *GetWeather*. Its limited number of classes, balanced label distribution, and well-defined semantic boundaries make it particularly suitable for analyzing how RNNs internalize and separate intent representations.

Applying the dynamical systems framework to this dataset reveals a robust spatial organization in the hidden state space across different model configurations [26]. Fig. 2 highlights various aspects of the geometry and evolution of hidden state trajectories. The hidden state space forms a low-dimensional manifold, within which each input sentence induces a trajectory defined by a sequence of hidden states, $\mathbf{h}_1, \ldots, \mathbf{h}_T$, as shown in Figs. 2(a), (b) and (c). These trajectories evolve from their initial positions toward distinct peripheral regions of the manifold, as shown in Figs. 2(d) and (e). These outer regions align with the direction vectors defined by the rows of the readout matrix, Figs. 2(e) and (f). This spatial alignment allows the network to reliable map trajectories to class-specific predictions.

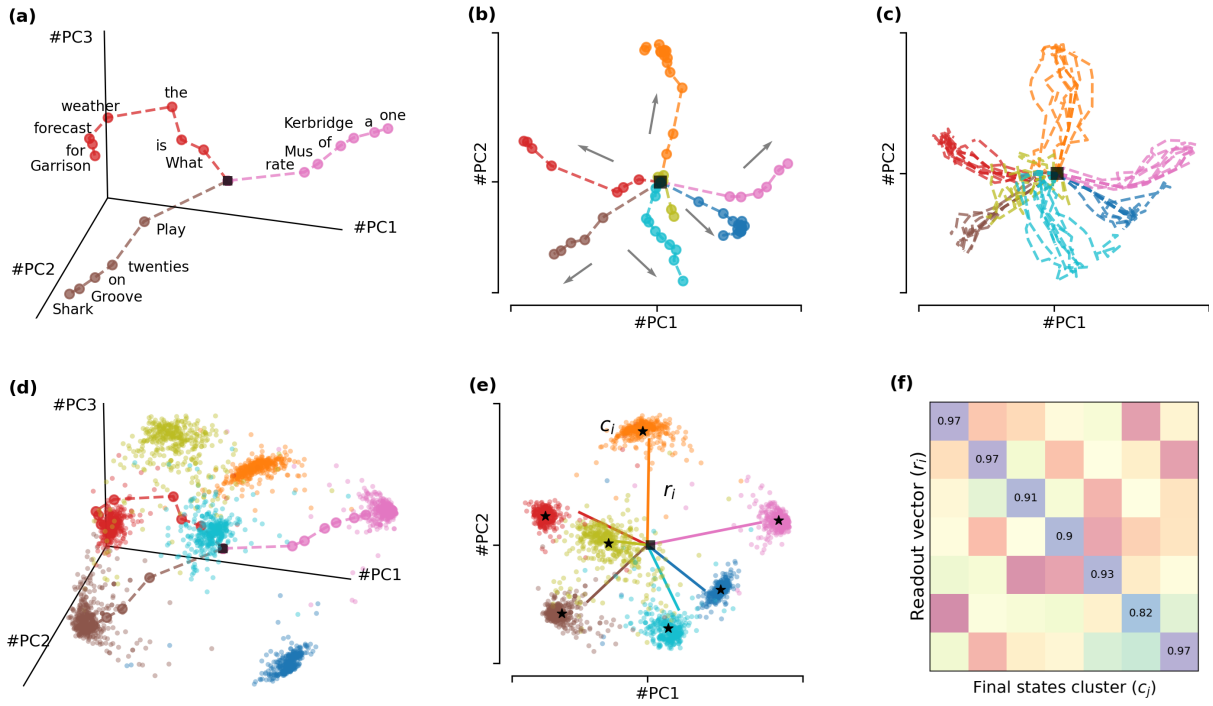



Fig. 2. State space visualizations of a GRU trained on the SNIPS dataset, projected onto the top principal components. **(a)** Example trajectories for three individual sentences. Each point represents a hidden state corresponding to a token in the sentence, a black square marks the initial hidden state. Dashed lines show transitions between states. **(b)** A representative trajectory for each intent, with arrows indicating the direction of movement. **(c)** Overlay of multiple trajectories sharing the same intent. **(d)** Three trajectories embedded in the 3D state space, superimposed on the final hidden states of all sentences. **(e)** Clusters of final states ($c_i$) with centroids marked as black stars. Readout vectors ($r_i$) from the output layer are color-coded to match their corresponding clusters . **(f)** Cosine similarities between $c_i$ and $r_i$. Only values above 0.5 are displayed [26].

These findings suggest that intent classification emerges from the underlying geometric organization of the hidden state space, offering valuable insights into how RNNs encode and separate intent categories. Beyond enhancing interpretability, the application of alternative analytical frameworks opens the door to new lines of inquiry that are difficult to pursue with conventional approaches. These include questions about trajectory convergence patterns and structural differences across classes or demographic groups, enabling more nuanced analyses of model bias and representational disparities.

## 4 Bias Analysis Based on State Space Dynamics

Bias in NLP has been analyzed from multiple perspectives [4]. In the context of pre-trained embeddings [21, 22], various bias measurement techniques [6, 18] and debiasing methods [5, 23] have been proposed. Many of these approaches rely on projecting into a low-dimensional vector space before measuring or debiasing [3]. Task-specific bias analysis have been conducted in sentiment analysis [10, 17], coreference resolution [25, 29], and machine translation [12].

We introduce a novel approach for analyzing bias and fairness in NLP by leveraging the sequential nature of language and applying reverse-engineering techniques from dynamical systems theory. This perspective provides new methodologies for detecting and mitigating bias. In our framework, sentence trajectories represent decision-making paths, offering intuitive visualizations of how user sentences are processed. By analyzing trajectory characteristics, final state distributions, and readout vector alignments across different sociodemographic and linguistic groups, we can identify structural effects in the state space that may lead to disparate error rates and misclassifications. Specifically, we focus on the task of intent detection.

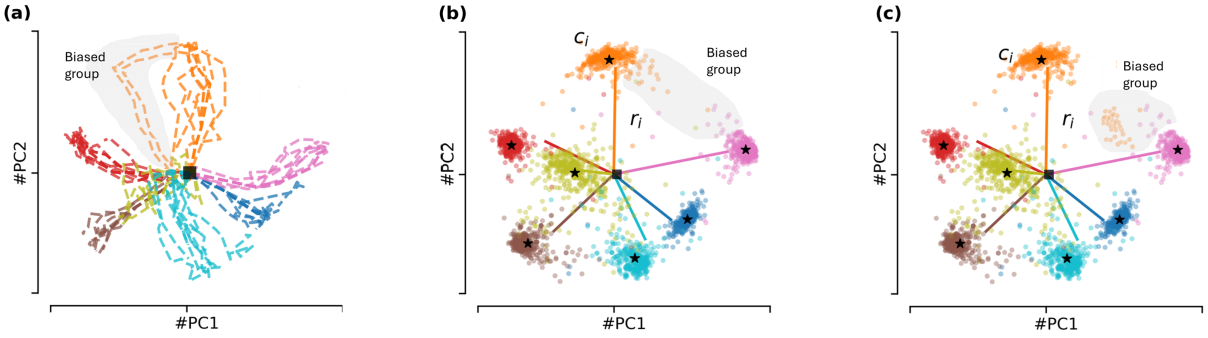



Fig. 3. Visualization of potential bias effects (highlighted as shaded areas) in the state space of an RNN trained for intent detection, projected onto the top principal components. **(a)** Sentences from the biased group reach the correct cluster but traverse trajectories that remain near the boundary of the basin of attraction, increasing the risk of misclassification during processing, even if they are ultimately classified correctly. **(b)** The final states of the biased group remain close to the correct attractor but do not integrate tightly with the core of the cluster, leading to lower confidence scores. **(c)** The final states of the biased group drift toward an incorrect attractor, resulting in misclassification.

To systematically assess bias, we propose analyzing three key aspects of the hidden state space dynamics. As input sentences evolve toward their corresponding attractor cluster, the resulting trajectories exhibit varying degrees of sparsity and organization. Figure 3 outlines three bias patterns that could emerge in underrepresented

demographic groups, emphasizing the interaction between trajectory structure and prediction reliability. As shown in Fig. 3(a), disparities in trajectory density and concentration within specific regions of the basin of attraction of each attractor can indicate the existence of bias. A second possible effect is depicted in Fig. 3(b), differences in cluster compactness and separability between groups might reflect bias, with more disperse state-space configurations being more susceptible to misclassification. Finally, Fig. 3(c) illustrates how the final states of a biased group could drift toward incorrect basins of attraction, aligning with unintended readout directions and potentially leading to systematic classification errors.

While the present work focuses on the qualitative characterization of bias patterns in the hidden state space, we take a step toward a quantitative framework by proposing a set of preliminary metrics designed to capture the three patterns outlined in Fig. 3. Trajectory density within an attractor basin could be estimated using kernel density estimation (KDE) applied to intermediate hidden states along the trajectory. This would enable group-level comparisons of how tightly or diffusely trajectories evolve over time, potentially revealing subtle disparities in how different inputs are processed within the model's internal dynamics. Similarly, the compactness and separability of final hidden states could be quantified using pairwise cosine distances within and between demographics groups, or alternatively via silhouette scores computed with cosine distance as the similarity metric, offering a geometric measure of how distinctly different groups are encoded. For the third pattern, alignment with readout directions could be approximated by measuring the cosine similarity between final hidden states and the corresponding row vectors of the readout matrix; systematic deviations in this alignment may indicate representational imbalances across groups. Although these metrics remain preliminary and unvalidated, they offer a concrete foundation for future work aimed at formalizing state-space-based fairness assessments and integrating them into model evaluation pipelines.

## 5 Conclusion and Future Work

In this work, we introduce a novel application of dynamical systems theory to the analysis of bias and fairness in NLP. Our approach frames the internal behavior of RNNs in terms of trajectory dynamics within the hidden state space, offering an interpretable framework to examine the presence of demographic or linguistic bias. We hypothesize that the presence of such biases can influence the geometric structure of these trajectories, leading to systematic representational differences. As a next step, we plan to extend this framework to multi-class classification tasks, where bias may lead to adverse effects. We aim to validate our approach using benchmark fairness datasets such as *Bias in Bios* [9] and *Mozilla Common Voice* [2], which provide rich demographic variability and real-world linguistic diversity. A key goal is to formalize the proposed state-space bias patterns into formal metrics, quantifying trajectory sparsity, compactness and separability of final representations, and alignment with readout directions. In parallel, a particularly challenging aspect lies in adapting these methods to more complex architectures, such as transformers.

## Acknowledgments

MAGN acknowledges the support by the European Union HORIZON-CL4-2021-HUMAN-01-01 under grant agreement 101070028 (REXASI-PRO) and by TED2021-129438B-I00 / AEI/10.13039/501100011033 / Unión Europea NextGenerationEU/ PRTR.

## References


[1] Kyle Aitken, Vinay V. Ramasesh, Ankush Garg, Yuan Cao, David Sussillo, and Niru Maheswaranathan. 2021. The geometry of integration in text classification RNNs. In *International Conference in Learning Representation (ICLR)*.

[2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, 4218–4222.

[3] Rajas Bansal. 2022. A Survey on Bias and Fairness in Natural Language Processing. arXiv:2204.09591.

[4] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 5454–5476.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, Vol. 29.

[6] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[7] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1724–1734.

[8] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips Voice Platform: an embedded Spoken Language Understanding system for private-by-design voice interfaces.

[9] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 120–128.

[10] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2019. Addressing Age-Related Bias in Sentiment Analysis. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*. 6146–6150.

[11] Jeffrey L. Elman. 1990. Finding Structure in Time. *Cognitive Science* 14, 2 (1990), 179–211.

[12] Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 147–154.

[13] Dilek Hakkani-Tür, Gokhan Tur, Asli Celikyilmaz, Yun-Nung Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-Domain Joint Semantic Frame Parsing Using Bi-Directional RNN-LSTM. In *Proc. InterSpeech*. 715–719.

[14] Robert Haschke and Jochen J. Steil. 2005. Input space bifurcation manifolds of recurrent neural networks. *Neurocomputing* 64 (2005), 25–38.

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[16] Hassan K. Khalil. 2013. *Nonlinear Systems*. Pearson.

[17] Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, 43–53.

[18] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 166–172.

[19] Bing Liu. 2015. *Sentiment Analysis and Opinion Mining*. Cambridge University Press.

[20] Niru Maheswaranathan, Alex Williams, Matthew Golub, Surya Ganguli, and David Sussillo. 2019. Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, Vol. 32. 15670–15679.

[21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *International Conference in Learning Representation 2018*.

[22] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 1532–1543.

[23] Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing Embeddings for Reduced Gender Bias in Text Classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 69–75.

[24] Libo Qin, Tianbao Xie, Wanxiang Che, and Ting Liu. 2021. A Survey on Spoken Language Understanding: Recent Advances and New Frontiers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. 4577–4584.

[25] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 8–14.

[26] Eduardo Sanchez-Karhunen, Jose F. Quesada-Moreno, and Miguel A. Gutiérrez-Naranjo. 2024. Interpretation of the Intent Detection Problem as Dynamics in a Low-Dimensional Space. In *Frontiers in Artificial Intelligence and Applications*. Vol. 392. 3693–3700.

[27] Steven H. Strogatz. 2015. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (2nd ed.). CRC Press.

[28] David Sussillo and Omri Barak. 2013. Opening the Black Box: Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks. *Neural Computation* 25, 3 (2013), 626–649.

[29] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2979–2989.