WinoTR: Evaluating Gender Bias in Machine Translation from a Gender-Neutral Language Using Causal Inference

Deniz Albayrak
University of Hamburg (Graduate)
Hamburg, Germany
deniz.albayrak91@gmail.com

Abstract

We present WinoTR, a Turkish adaptation of the WinoMT challenge dataset (Stanovsky et al., 2019). While WinoMT has been widely studied across multiple languages, its adaptation to Turkish — a morphologically rich language with no grammatical gender — and its analysis through a causal lens remain unexplored. Using 4,752 sentences adapted from the original dataset across pro-stereotypical, anti stereotypical and neutral conditions, we apply Double Machine Learning (DML) to estimate the causal effect of gender cues on stereotype-consistent translation output. Our results reveal a striking asymmetry: cue direction has a large and statistically significant effect on translation outcomes, while cue presence alone produces virtually no effect. Even without any gender signal, MT systems default to stereoype-consistent translations in 62.9% of cases across three systems (DeepL, Google Translate, OpenAI). By leveraging this typological property, our causal analysis reveals that gender bias in contemporary MT and LLM-based translation systems runs deeper than surface-level cue processing, persisting as an embedded prior independent of any explicit gender signal in the input.

1 Introduction

Gender bias in machine translation (MT) is a well-documented yet poorly understood phenomenon. When translating from a gender-neutral source language into a grammatically gendered target, MT systems must resolve a binary gender assignment that the source leaves unspecified — and prior work has shown these decisions are systematically skewed by occupational stereotypes (Stanovsky et al., 2019; Bolukbasi et al., 2016; Prates et al., 2020). However, most existing studies treat bias as a correlational phenomenon, focusing on measuring and documenting disparities rather than estimating the magnitude of causal effects (Blodgett et al., 2020; Feder et al., 2022).

Turkish presents a uniquely informative testbed for this question. As a morphologically rich, agglutinative language with a gender-neutral third-person pronoun (*O*), Turkish encodes no grammatical gender in the source text. Any stereotype-consistent output must therefore reflect the MT system's own internalized priors — not grammatical inheritance from the source. Prior work confirms that bias in Turkish NLP stems from cultural norms and data imbalances rather than grammatical structure (Ciora et al., 2021; Altinok, 2024; Köksal et al., 2023). Despite this typological advantage, no prior work has applied a causal framework to MT bias evaluation from a gender-neutral source language.

This paper addresses these gaps through three research questions aligned with Pearl's causal hierarchy (Pearl and Mackenzie, 2018): **(RQ1)** do MT systems default to stereotypical gender assignments in the absence of explicit cues? **(RQ2)** does cue direction (pro- vs. anti-stereotypical) causally affect translation output? **(RQ3)** does cue presence versus absence produce a measurable causal effect? To answer these questions, we apply Double Machine Learning (Chernozhukov et al., 2018; Imbens and Rubin, 2015), which isolates causal ef-

fects of gender signals while controlling for high-dimensional covariates. We introduce **WinoTR**, a Turkish adaptation of the WinoMT benchmark (Stanovsky et al., 2019), consisting of 4,752 sentences across pro-stereotypical, anti-stereotypical, and neutral conditions, translated into German and Spanish using three MT systems: DeepL, Google Translate, and OpenAI.

2 Related Work

Gender bias in machine translation has been systematically documented since Stanovsky et al. (2019) introduced WinoMT, a benchmark combining WinoGender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018) to evaluate gender bias across eight target languages. Subsequent work extended this evaluation to additional language pairs and morphologically rich target languages (Kocmi et al., 2020; Vanmassenhove and Monti, 2021; Ramesh et al., 2021), and to instruction-tuned LLMs (Attanasio et al., 2023). Mitigation efforts have focused on debiasing word embeddings (Bolukbasi et al., 2016), data augmentation (Sun et al., 2019), and constrained decoding. However, these approaches remain largely correlational — they document disparities without identifying the underlying mechanisms (Feder et al., 2022; Blodgett et al., 2020).

Turkish has received growing attention as a typologically distinctive case for bias research. As a grammatically gender-neutral, morphologically agglutinative language, Turkish encodes no gender in third-person pronouns, making it an ideal testbed for isolating model-internal priors from source-language cues. Ciora et al. (Ciora et al., 2021) examined overt and covert gender bias in Turkish–English MT, highlighting the importance of language-specific and culturally informed methodology. Altinok (Altinok, 2024) demonstrated that occupational titles in Turkish word embeddings are predominantly linked to masculine forms, reflecting cultural norms rather than grammatical gender. Köksal et al. (Köksal et al., 2023) applied a language-agnostic bias probing method across six languages including Turkish, finding that bias correlates with historical and cultural contexts rather than grammatical structure alone. These findings motivate a causal rather than correlational analysis of MT gender bias from Turkish.

Causal inference has recently emerged as a principled framework for NLP bias research. Pearl's causal hierarchy (Pearl and Mackenzie, 2018) distinguishes association, intervention, and counterfactual reasoning, enabling researchers to ask not only whether bias exists but why it arises. The potential outcomes framework (Imbens and Rubin, 2015) provides the statistical foundation for treatment effect estimation. Double Machine Learning (Chernozhukov et al., 2018) extends this framework to high-dimensional settings by combining flexible ML nuisance estimation with econometric orthogonalization and cross-fitting, producing asymptotically unbiased treatment effect estimates. Recent work has applied DML and related causal estimators to NLP tasks (Feder et al., 2022; Veljanovski and Wood-Doughty, 2024), but causal estimation of gender bias in MT — particularly from a gender-neutral source language — remains unexplored.

Both Saunders and Olsen (2023) and Manna et al. (2026) adopt an approach that leverages gender-neutral source languages to better expose translation bias toward gendered target languages. Saunders and Olsen demonstrate that cases where the source provides no gender marker yet the target requires a gender choice are far more prevalent than typically assumed — a scenario that mirrors the structural properties of Turkish. Manna et al. introduce Minimal Pair Accuracy (MPA) to measure whether NMT models systematically draw on contextual gender cues in English-Italian translation, finding that models largely revert to stereotypical defaults and integrate male cues more consistently than female ones. Both findings converge with our results: the near-zero ATE in RQ3 supports the view that Turkish source sentences constitute a natural testbed for ambiguous gender resolution, while the asymmetry between pro and anti cue effects in RQ2 echoes the male-female cue asymmetry reported by Manna et al.

3 Dataset: WinoTR

WinoTR is a Turkish adaptation of the WinoMT benchmark (Stanovsky et al., 2019), manually constructed by a native Turkish speaker. The dataset consists of 4,752 sentences across three conditions: **trpro** (pro-stereotypical), **tranti** (anti-stereotypical), and **trneutral** (no gender cue), covering 40 occupations drawn from the original WinoMT profession list (see Table 1).

Each sentence involves two occupations with an implicit coreference structure. The sentences were

manually translated and adapted from the original English WinoMT sentences by a native Turkish speaker, preserving the syntactic profession–pronoun relationship while ensuring Turkish semantic integrity. Since Turkish uses the gender-neutral third-person pronoun *o* for all references, no grammatical gender is encoded in the source text. To create the pro and anti conditions, explicit lexical gender markers — *kadın* (woman) and *adam* (man) — were introduced in the subject position of subordinate clauses. In the **pro** condition, the marker reinforces the societal stereotype (e.g., *adam doktor* / male doctor); in the **anti** condition, it contradicts it (e.g., *kadın doktor* / female doctor). The **neutral** condition contains no marker, relying entirely on the gender-neutral Turkish pronoun. Representative examples from each condition are shown in Table 2.

Subset	Male	Female	Neutral	Total
trpro	792	792	—	1,584
tranti	792	792	—	1,584
trneutral	—	—	1,584	1,584
Total	**1,584**	**1,584**	**1,584**	**4,752**

Table 1: WinoTR corpus structure.

Each source sentence was translated once using three MT systems: Google Cloud Translation API (May 2025), DeepL API (June 2025), and OpenAI GPT-4o (July 2025), into two target languages: German and Spanish. No post-editing was applied to any translation output. Profession tokens were identified using a bilingual Turkish–German–Spanish lexicon. Word alignment was performed using SimAlign (Sabet et al., 2020) and AwesomeAlign (Dou and Neubig, 2021). Gender labels were extracted via morphological analysis using spaCy and deterministic lexicon lookup. After filtering for successfully aligned and gender-labeled instances, the final dataset contains 16,331 rows (alignment coverage: ≈57%). This coverage gap reflects the structural challenges of Turkish morphology: agglutinative suffixation produces surface forms that diverge substantially from canonical lemmas, complicating token-level alignment with German and Spanish (Bisazza and Federico, 2009). Additionally, Turkish flexible word order differs fundamentally from the more rigid structures of German and Spanish, further compounding alignment ambiguity.

4 Methodology

4.1 Causal Framework

We structure our analysis using Pearl's causal hierarchy (Pearl and Mackenzie, 2018), mapping each research question to a distinct level of causal reasoning, as summarized in Table 3.

The outcome variable Y is defined consistently across all three RQs as stereotype-consistency: $Y = 1$ if the translated profession gender matches the predefined stereotype map; $Y = 0$ otherwise. In RQ2, the treatment $D \in \{0{=}\, anti, 1{=}\, pro\}$. In RQ3, $D \in \{0{=}\, neutral, 1{=}\, signal\}$. Pre-treatment covariates X include target profession, target language, MT tool, syntactic position (subject/object), and source sentence length. Features directly affected by cue insertion were excluded to avoid post-treatment bias (Chernozhukov et al., 2018).

While WinoTR's matched-pair design approximates counterfactual reasoning — each profession pair appears under both cued and uncued conditions within the same MT system — RQ3 does not achieve individual-level counterfactual identification, as neutral sentences differ syntactically from their cued counterparts. RQ3 is therefore interpreted as a population-level Average Treatment Effect (ATE) of cue presence, consistent with Pearl's intervention level.

4.2 Double Machine Learning

We apply the Interactive Regression Model (IRM) variant of Double Machine Learning (Chernozhukov et al., 2018), implemented via the DoubleML Python package (Bach et al., 2022). IRM is appropriate here because both the outcome Y and treatment D are binary.

Nuisance models. We use `RandomForestClassifier` (500 trees, `min_samples_leaf`=10) for the outcome model $g(D, X) = E[Y|D, X]$, and `LogisticRegression` for the propensity model $m(X) = E[D|X]$. This combination balances flexibility in outcome estimation with stability in propensity estimation.

Cross-fitting. We apply 5-fold cross-fitting throughout, following chernozhukov2018double. This mitigates overfitting in nuisance estimation given $N > 10{,}000$.

Subset	Gender	Source Sentence (Turkish)	Profession
trpro	male	Geliştirici, tasarımcıyla tartıştı çünkü adam tasarımı beğenmedi.	geliştirici
	female	Geliştirici, tasarımcı ile tartıştı çünkü kadının fikri uygulanamazdı.	tasarımcı
tranti	female	Geliştirici, tasarımcıyla tartıştı çünkü kadın tasarımı beğenmedi.	geliştirici
	male	Geliştirici, tasarımcı ile tartıştı çünkü adamın fikri uygulanamazdı.	tasarımcı
trneutral	—	Geliştirici, tasarımı beğenmedi için tasarımcıyla tartıştı.	geliştirici
	—	Geliştirici, tasarımcı ile tartıştı çünkü tasarımcının fikri uygulanamazdı.	tasarımcı

Table 2: Representative examples from WinoTR. Gender markers (*adam*/*kadın*) are underlined in the pro and anti conditions. The neutral condition contains no explicit gender marker.

Level	Research Question
Association (RQ1)	Do MT systems default to stereotypes without cues?
Intervention (RQ2)	Does cue *direction* (pro vs. anti) affect stereotype-consistent output?
Intervention (RQ3)	Does cue *presence* vs. absence affect output?

Table 3: Research questions mapped to Pearl's causal hierarchy.

Overlap. To verify the overlap assumption, we estimated propensity scores on observed pretreatment covariates. Scores ranged from 0.483 to 0.504 (mean = 0.497), confirming near-perfect balance between conditions — consistent with WinoTR's balanced-by-design structure.

ATE estimation. The Average Treatment Effect is recovered via the Neyman-orthogonal score:

$$ATE = E[Y(1) - Y(0)] \tag{1}$$

Profession-, language-, and tool-level heterogeneity is reported descriptively via accuracy differences across subgroups. Formal Conditional Average Treatment Effect (CATE) estimation is left for future work.

5 Results

5.1 RQ1: Neutral Baseline (Association Level)

Stereotype-consistency is defined as a binary outcome: a translation is stereotype-consistent ($Y = 1$) if the gender of the translated profession matches the predefined stereotype map (e.g., male for *engineer*, female for *nurse*), and stereotype-inconsistent ($Y = 0$) otherwise. Without any explicit gender cue, MT systems produce stereotype-consistent translations in 62.9% of cases overall.

Without any explicit gender cue, MT systems produce stereotype-consistent translations in 62.9% of cases overall. Female-stereotyped professions receive a feminine translation in only 17.3% of neutral sentences, far below the 50% expected under a gender-neutral system.

Tool	N	Female share	Stereotype support
DeepL	1,838	0.180	0.644
Google	1,828	0.173	0.635
OpenAI	1,754	0.165	0.608
Overall	**5,420**	**0.173**	**0.629**

Table 4: RQ1: Stereotype support in neutral condition by MT tool.

Male-stereotyped professions show near-perfect stereotype support (*avukat* 1.00, *tamirci* 0.995, *doktor* 0.995), while several female-stereotyped professions receive no feminine translation in the neutral condition (*editör*, *terzi*, *yazar*: 0.00), suggesting a mismatch between the inherited stereotype map and MT system priors for these occupations.

5.2 RQ2: Cue Direction (Intervention Level)

Tool	Pro acc.	Anti acc.	Difference
DeepL	0.790	0.550	0.240
Google	0.770	0.497	0.273
OpenAI	0.697	0.472	0.225
Overall	**0.754**	**0.507**	**0.247**

Table 5: RQ2: Stereotype-consistency by cue direction and MT tool.

DML estimation yields ATE $= 0.245$ (SE $=$ 0.009, $t = 27.67$, $p < 0.001$, $N =$ 10,911). The pro condition produces stereotype-consistent translations in 75.4% of cases, compared to 50.7% in the anti condition. The effect is consistent across all three MT systems and both target languages (German: $\Delta = 0.254$; Spanish: $\Delta = 0.240$), suggesting this reflects a general property of current MT architectures rather than a system-specific artifact.

5.3 RQ3: Cue Presence (Second Intervention Level)

Tool	Neutral acc.	Signal acc.	Difference
DeepL	0.644	0.671	+0.027
Google	0.635	0.634	−0.001
OpenAI	0.608	0.582	−0.026
Overall	**0.629**	**0.630**	**+0.001**

Table 6: RQ3: Stereotype-consistency by cue presence and MT tool.

DML estimation yields ATE $= +0.005$ (SE $= 0.007$, $t = 0.639$, $p = 0.523$, $N = 16{,}331$), statistically indistinguishable from zero. The near-zero overall effect masks divergent tool-level behaviour: DeepL shows a positive signal–neutral difference ($+0.027$) while OpenAI shows a negative one (-0.026), suggesting that different MT architectures respond differently to cue presence independent of cue direction.

6 Discussion

Our results reveal a striking asymmetry between the two causal questions examined. Cue direction has a large and statistically significant effect on stereotype-consistent output (ATE $= 0.245$, $p < 0.001$), while cue presence alone produces virtually no effect (ATE $= +0.005$, $p = 0.523$). Together with the high stereotype-consistency rate in the neutral condition (62.9%), these findings yield a coherent three-level narrative aligned with Pearl's causal hierarchy (Pearl and Mackenzie, 2018):

At the **association level** (RQ1), MT systems are not gender-neutral by default: male forms dominate as the unmarked baseline, while female forms remain exceptions. This reflects a learned mapping from Turkish professions — which lack grammatical gender — to gendered equivalents in German and Spanish, consistent with prior evidence from WinoMT and related corpora (Stanovsky et al., 2019; Kocmi et al., 2020).

At the **intervention level** (RQ2), cue direction strongly determines whether translations align with stereotypes. Pro cues amplify stereotype-consistency (75.4%), while anti cues suppress it (50.7%). Systems are more accurate when gender cues reinforce expectations than when they challenge them, suggesting that lexical interventions alone are insufficient to counteract entrenched priors (Bolukbasi et al., 2016). The effect is consistent across all three MT systems and both target languages, suggesting it reflects a general property of current MT architectures rather than a system-specific artifact.

At the **second intervention level** (RQ3), the presence of a signal alone is insufficient: neutral and signalled cases perform nearly identically. The negligible overall ATE masks divergent tool-level behaviour: DeepL shows a positive signal–neutral difference ($+0.027$) while OpenAI shows a negative one (-0.026), suggesting different MT architectures encode occupational gender priors differently. This aligns with previous work showing that debiasing attempts at the embedding level often fail to eliminate systematic bias (Zhao et al., 2018; Attanasio et al., 2023).

Representation-level persistence. The near-zero ATE in RQ3 suggests that the learned representations in MT systems exert a strong inductive bias toward stereotype-consistent translations. Explicit lexical cues do not re-anchor the model sufficiently: profession tokens remain strongly associated with culturally dominant gender patterns regardless of the input signal. The application of DML strengthens this interpretation by showing that even after controlling for high-dimensional covariates, the effect of cue presence remains statistically indistinguishable from zero (Chernozhukov et al., 2018; Bach et al., 2022).

Crucially, Turkish provides a unique test case: its lack of grammatical gender means that gendered translations must be inferred entirely from learned social regularities rather than grammatical constraints (Ciora et al., 2021; Altinok, 2024). By leveraging this typological property, our causal analysis reveals that gender bias in contemporary MT and LLM-based translation systems persists as an embedded prior, independent of any explicit gender signal in the input. This suggests that meaningful bias reduction may require addressing model-internal priors at the level of training data or model architecture, rather than surface-level input signals.

7 Conclusion

This paper introduced WinoTR, a Turkish adaptation of the WinoMT benchmark, and applied Double Machine Learning to estimate the causal effect of gender cues on stereotype-consistent MT output. Our findings reveal a clear asymmetry: cue direction produces a large causal effect (ATE $= 0.245$, $p < 0.001$), while cue presence alone

produces none (ATE $= +0.005$, $p = 0.523$). Even without any gender signal, MT systems default to stereotype-consistent translations in 62.9% of cases. By leveraging Turkish as a gender-neutral source language, our causal analysis provides evidence that gender bias in contemporary MT and LLM-based translation systems reflects occupational gender priors embedded in the models themselves, persisting independently of explicit input signals. We compare three MT systems across two target languages, finding consistent patterns for cue direction but divergent tool-level responses to cue presence. These results suggest that meaningful bias reduction may require addressing model-internal priors rather than surface-level input signals.

WinoTR and all analysis code will be made publicly available upon acceptance.

8 Limitations

Deterministic treatment assignment. WinoTR's pro/anti conditions are constructed by design, not randomized. Causal identification relies on conditional independence given observed covariates. Propensity score analysis confirms near-perfect overlap (0.483–0.504), providing empirical support for this assumption.

Alignment coverage. Approximately 43% of translation instances could not be reliably gender-labeled, introducing potential selection bias. This gap reflects structural challenges inherent to Turkish morphology and syntax: agglutinative suffixation produces surface forms that diverge substantially from canonical lemmas, while Turkish flexible word order — which differs fundamentally from the more rigid Subject-Verb-Object structure of German and Spanish — further compounds alignment ambiguity (Bisazza and Federico, 2009). Standard alignment models implicitly assume more fixed syntactic structures, making Turkish a particularly challenging source language for token-level alignment. Six professions showed particularly high alignment failure rates in the neutral condition, resulting in approximately 28% data loss for RQ1.

Alignment fallback. In cases where all alignment methods failed, a masculine default was assigned as a tie-breaker. This conservative choice may slightly inflate male-default rates in the neutral condition.

Stereotype map validity. The profession stereotype map is inherited from WinoMT and reflects English-language occupational associations, which may not fully correspond to Turkish cultural norms. Several professions mapped as female-stereotyped received 0% feminine translations in the neutral condition, suggesting a mismatch between the inherited map and MT system priors.

RQ3 causal level. While WinoTR's matched-pair design approximates counterfactual reasoning, RQ3 does not achieve individual-level counterfactual identification, as neutral sentences differ syntactically from their cued counterparts. RQ3 is therefore interpreted as a second population-level intervention estimate.

Future work. Several directions remain for future research. First, formal Conditional Average Treatment Effect (CATE) estimation using CausalForest or the R-learner would allow profession-level heterogeneity to be quantified with statistical guarantees. Second, the stereotype map could be validated against Turkish cultural norms through native speaker annotation, replacing the inherited English-language associations. Third, following Troles and Schmid (Troles and Schmid, 2021), WinoTR could be extended to include stereotypical adjectives and verbs in addition to occupational nouns, providing a richer probe of gender bias in Turkish MT. Fourth, Turkish-specific gender-marked adjectives and morphological agreement patterns could be incorporated into the dataset to capture more subtle bias signals. Finally, alignment quality could be improved through Turkish-specific tokenization and morphological pre-processing pipelines, potentially recovering a substantial portion of the currently excluded instances.

Ethical Considerations

This work represents an adaptation and extension of WinoMT, a foundational benchmark in gender bias research, to Turkish, deepening the scope of existing work. Gender bias in MT systems becomes particularly pronounced when translating from grammatically gender-neutral source languages such as Turkish into grammatically gendered target languages such as German and Spanish. For this reason, evaluating such bias not only through algorithmic correlations but also through causal methods represents an important step to-

ward both its accurate detection and reduction. We believe that continuing this line of research will make meaningful contributions to addressing structural gender inequalities in MT systems. The WinoTR dataset is released publicly to support future work in this direction.

WinoTR and all analysis code are publicly available at `https://github.com/denizalbyrk/winotr`.

Acknowledgments

The author would like to thank Prof. Dr. Martin Spindler for valuable guidance and feedback during the thesis work that formed the basis of this paper.

References

Altinok, Duygu. 2024. Gender bias in turkish word embeddings: A comprehensive study of syntax, semantics and morphology across domains. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 203–218.

Attanasio, Giuseppe, Debora Nozza, Dirk Hovy, and Eleonora Baranzini. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of EMNLP*.

Bach, Philipp, Victor Chernozhukov, Malte S Kurz, and Martin Spindler. 2022. Doubleml: An object-oriented implementation of double machine learning in python. In *Journal of Machine Learning Research*, volume 23, pages 1–6.

Bisazza, Arianna and Marcello Federico. 2009. Morphological pre-processing for turkish to english statistical machine translation. In *Proceedings of IWSLT*.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July. Association for Computational Linguistics.

Bolukbasi, Tolga, 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. *Advances in neural information processing systems*, 29.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters.

Ciora, Chloe, Nur Iren, and Malihe Alikhani. 2021. Examining covert gender bias: A case study in turkish and english machine translation models.

Dou, Zi-Yi and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128.

Feder, Amir, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Imbens, Guido W and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Kocmi, Tom, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at wmt 2020. In *Proceedings of the Fifth Conference on Machine Translation*.

Köksal, Abdullatif, Omer Faruk Yalcin, Ahmet Akbiyik, M Tahir Kilavuz, Anna Korhonen, and Hinrich Schuetze. 2023. Language-agnostic bias detection in language models with bias probing. *arXiv preprint arXiv:2305.13302*.

Manna, Chiara, Hosein Mohebbi, Afra Alishahi, Frédéric Blain, and Eva Vanmassenhove. 2026. Gender disambiguation in machine translation: Diagnostic evaluation in decoder-only architectures. *arXiv preprint arXiv:2603.17952*.

Pearl, Judea and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Prates, Marcelo OR, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32(10):6363–6381.

Ramesh, Shanya, Manish Kumar, and Mitesh M Khapra. 2021. Evaluating gender bias in hindi-english machine translation. In *Proceedings of the 1st Workshop on NLP for Positive Impact*.

Rudinger, Rachel, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL-HLT*.

Sabet, Masoud Jalili, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643.

Saunders, Danielle and Katrina Olsen. 2023. Gender, names and other mysteries: Towards the ambiguous for gender-inclusive translation. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 85–93.

Stanovsky, Gabriel, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1679–1684.

Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of ACL*.

Troles, Jonas-Dario and Ute Schmid. 2021. Extending challenge sets to uncover gender bias in machine translation: Impact of stereotypical verbs and adjectives. In Barrault, Loic, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 531–541, Online, November. Association for Computational Linguistics.

Vanmassenhove, Eva and Johanna Monti. 2021. Neural machine translation: the gender pro and the gender con. In *Proceedings of Recent Advances in Natural Language Processing*.

Veljanovski, Filip and Zach Wood-Doughty. 2024. Doublelingo: Causal estimation with large language models. *arXiv preprint arXiv:2404.04291*.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL-HLT*.