Barclay, P. J. and A. Sami (Apr. 2024)	Investigating Markers and Drivers of Gender Bias in Machine Translations	Investigates implicit gender bias in LLMs using back-translation with DeepL Translates 56 software engineering tasks starting with "she" into genderless languages (Finnish, Indonesian, Estonian, Turkish, Hungarian) and back to English Examines pronoun choices in back-translated texts Extends prior research by Treude and Hata	Comparison of results across five intermediate languages Proposed novel metric for variation in gender across repeated translations: coefficient of unalikeability (UCA) Investigated sentence features that drive bias, especially main verb Compared results from three time-lapsed datasets to test reproducibility	The variation in pronoun selection (quantified by UCA) indicates the language model's uncertainty or hesitancy in implying a particular gender, mirroring human usage of gender-neutral terms. The differing levels of pronoun variation in generated texts for certain tasks have the potential to subtly reinforce gender stereotypes over repeated exposure. Future research should involve more sentences, different translation APIs, longer phrases to capture more context-dependent bias, and further investigation into the correlation between verbs and pronoun variation. The UCA metric allows probing biases without making assumptions about what constitutes a biased formulation.	Different intermediate languages display varying patterns of pronoun use, falling into three groups: Finnish and Estonian (frequent 'he', moderate 'he/she', few missing pronouns), Hungarian and Turkish (many missing pronouns, greater 'you' use), and Indonesian (almost exclusive use of 'he').
Cho, W. I. et al. (2019)	On Measuring Gender Bias in Translation of Gender-neutral Pronouns	Focuses on Korean-to-English translation Korean has gender-neutral pronouns like "걔 (kyay)"	First attempt to evaluate gender bias in KR-EN translation for sentiment words and occupations Constructed Equity Evaluation Corpus (EEC) Introduced Translation Gender Bias Index (TGBI) to compare MT systems	Gender bias in machine translation (MT) systems, especially in the translation of gender-neutral pronouns, is not thoroughly investigated for cross-lingual tasks and can perpetuate real-world prejudice.	For sentences where gender determination is not explicitly provided by context, translation systems are recommended to use each gender equally or neutral pronouns if available, to avoid hasty guesses. Occupation translations were found to be more biased than other categories across all systems.
Godsil, R. D. et al. (2016)	The Effects of Gender Roles, Implicit Bias, and Stereotype Threat on the Lives of Women and Girls	Reviews social science research on gender bias, implicit bias, and stereotype threat Uses intersectional lens to assess impacts on academic and professional outcomes for women Notes disparities result from structural discrimination and social stereotypes, not talent	Evidence-based strategies to override bias: Increase diversity / critical mass: three or more women in professional settings improves governance, innovation, sense of belonging In-group peers: networking and peer mentoring fosters community Visible experts: showcasing women in underrepresented fields helps newcomers identify with success Provide effective task strategies for stereotype threat situations	The notion of gender has expanded beyond the binary, and the specific challenges faced by gender nonconforming, transgender, and LBQ individuals warrant separate, dedicated reports.	
Kappl, M. (2025)	Are All Spanish Doctors Male? Evaluating Gender Bias in German Machine Translation	Large-scale evaluation of five commercial MT systems (Google Translate, Microsoft Translator, Amazon Translate, DeepL, SYSTRAN) and GPT-4o-mini Evaluated German to seven target languages with gendered grammar: French, Italian, Spanish, Ukrainian, Russian, Arabic, Hebrew Pipeline: translation, prediction using word alignment and morphological analysis, metric calculation (accuracy, gender-based F1 gaps, stereotype-based performance gaps)	Introduced WinoMTDE: German gender bias evaluation set 288 German sentences based on Winograd schema Balanced gender subjects and stereotype alignment Currently only binary-gendered terms	The WinoMTDE dataset is relatively small (288 sentences), limiting the scope of bias assessment. Stereotype annotations were based on a single person and German labor statistics, potentially introducing bias, especially for ambiguous job titles. The dataset's exclusion of non-binary pronouns and neutral job titles restricts the analysis to a binary gender perspective and overlooks broader gender biases. Certain biases, like semantic derogation (e.g., "teacher" translating to gendered terms), remain unaddressed.	Problem stems from systematic bias within model and training data, not source language ambiguity. The results reveal persistent gender bias in most models across the tested languages. GPT-4o-mini generally outperformed traditional MT systems in terms of accuracy. The study visualizes predictions for occupation groups, showing how translated gender distribution often does not align with real-world distributions, highlighting biases.
Lardelli, M., G. Attanasio, and A. Lauscher (2024)	Building Bridges: A Dataset for Evaluating Gender-Fair Machine Translation into German	Creation of novel resources and MT system evaluation Gender-fair language (GFL) extremely rare, context does not significantly improve it Automatic detection in zero-shot settings is very challenging	Gender-Fair German Dictionary: lists gender-neutral and inclusive German terms with English translations Multi-domain test data from Wikipedia and Europarl Benchmarking GFL in eight translation systems (Google Translate, DeepL, GPT-3.5, GPT-4, NLLB, OPUS MT, Flan-T5, Llama 2)	The study focuses on a single language pair and direction (English to German), a relatively small number of seed nouns and sampled sentences, and deliberately focuses on sentences where the entity's gender is ambiguous or mixed, discarding cases where it is disambiguated.	Research on gender-fair MT is scarce, particularly for German, with existing studies covering only a limited number of languages, scenarios, and domains. Linguistic forms influence mental representation of gender identities, making gender equality in language a crucial goal. German GFL strategies explored: gender-neutral rewording using passive constructions, indefinite pronouns, gender-neutral terms, or participles instead of gendered nouns; gender-inclusive characters using symbols like , :, or _ to combine masculine and feminine forms (e.g., "derdie Leserin"). Words-in-isolation: all tested models demonstrate a heavy bias towards masculine forms (93–96%). Feminine forms are used seldom (2–4%), mainly for stereotypically female professions. Gender-neutral and gender-inclusive forms are rarer (0–2%), appearing mainly for already common gender-neutral words. Words-in-context (Europarl and
Prates, M. O. R., P. H. C. Avelar, and L. Lamb (2019)	Assessing Gender Bias in Machine Trans- lation – A Case Study with Google Translate	Sentence templates used: e.g., "ő egy ápolónő" (Hungarian for "s/he is a nurse") Lexical focus on job positions (from US Bureau of Labor Statistics) and 21 adjectives to explore bias beyond occupations Experiments conducted with 12 gender-neutral languages from diverse families (Hungarian, Finnish, Estonian, Japanese, Chinese) Korean and Nepali initially considered but omitted due to technical issues	Found male pronoun dominance in MT STEM fields consistently defaulted to male Some languages like Basque favored neutral pronouns Adjectives also showed bias (e.g., "Shy" more female, "Guilty" more male) Bias could not be explained by workplace demographics alone		The paper's findings, published as a preprint, received significant media coverage. On December 6, 2018, Google changed its policy to present both feminine and masculine official translations for ambiguous queries, acknowledging their model inadvertently replicated gender biases. The research highlights that gender bias is a statistical phenomenon independent of proprietary tools, suggesting MT engineers must address training data and implement solutions after training rather than relying on scarce unbiased texts. The study concludes unbiased results can be obtained with relatively low effort and marginal performance cost using existing debiasing algorithms.
Rescigno, A. A. and J. Monti (2023)	Gender Bias in Machine Translation: A Statistical Evaluation of Google Translate and DeepL for English, Italian and German	Statistical approach using MT-GenEval dataset Single sentence translation, repeated with context Objectives: detect gender stereotypes in MT systems (Google Translate, DeepL) for English-German and English-Italian Examine if extended context mitigates bias for ambiguous referents	Translation systems show masculine default bias Google Translate and DeepL biased toward masculine outputs Context improves performance, especially for DeepL Contextual errors occur but infrequently	The study suggests the need for more comprehensive evaluations with wider language combinations and more targeted, balanced datasets (e.g., GATE) in future work.	"There is currently no tool to notify them about it" → no detection tool. MT systems still show a strong tendency to default to the masculine gender. Adding context generally improves results but can occasionally lead to erroneous disambiguation.

Savoldi, B., J. Bastings, et al. (2025)	A Decade of Gender Bias in Machine Translation	Comprehensive review synthesising previous research, identifies field limitations, highlights findings and future directions	Empirical methods: Translating gender-neutral sentences and analyzing pronoun frequency Challenge sets with automatic metrics (WinoMT, MT-GenEval) Human-centered quantitative assessment of MT bias	English-centric focus: much of the existing research is overwhelmingly English-centric, often with English as the source language and another Western language as the target, creating a "winner-takes-all" scenario where well-supported languages receive most attention, risking perpetuation of anglocentric biases and overlooking cultural, linguistic, and societal differences. Lack of human engagement: there is a severe lack of direct human involvement in MT gender studies. Most human evaluations are model-centric, supporting structured assessments of model behaviour rather than exploring feedback and experiences of impacted user groups. This gap limits understanding of real-world harms.	Growth in research but gaps remain: increase in papers on gender bias in MT from 2019–2023. Significant gaps: overemphasis on English-centric approaches and tendency to treat bias as purely technical, disregarding social and ethical components. Limited contextual understanding: most studies focus on sentence-level translation, despite gender often requiring broader context. Binary gender focus with emerging inclusivity: majority of studies treat gender as binary, though research increasingly accounts for non-binary identities. Diversity of mitigation strategies with no "clear winner": half of reviewed papers propose strategies (data curation, fine-tuning, inference-time approaches, post-processing rewrites), often modular and scenario-specific.
Savoldi, B., S. Papi, et al. (2024)	What the Harm? Quantifying the Tangible Impact of Gender Bias in Machine Translation with a Human-centered Study	Quantifies gender bias impact in MT via human-centered study Measures effort and cost to correct gender-biased MT outputs Simulated post-editing task: participants corrected outputs for feminine and masculine references	Gender bias in MT leads to higher effort and cost for feminine translations	Study is limited to binary gender categories, acknowledging that this does not imply a binary view of gender identity. This choice was made for controlled experimental conditions, as non-binary and neutral expressions are not yet standardized and would introduce confounds related to participants' familiarity and cognitive load.	
Sczesny, S., M. Formanowicz, and F. Moser (2016)	Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination?	Discusses implementation and effectiveness of GFL policies Focuses on influence on mental representation of gender and reducing stereotyping	GFL more accepted with regulations and frequent use	More research is needed to evaluate the effectiveness of language-related policies and provide evidence-based rationale for policy-making, especially across different grammatical languages and stages of GFL implementation. Challenges and effectiveness of GFL: spontaneous use remains infrequent despite guidelines. Factors influencing use include deliberate processes (attitudes, intentions) and habitual processes (repetition of past behaviour), with context also playing a role (e.g., official texts vs informal communication). Attitudes and intentions toward GFL are only moderately favourable; future research should identify crucial factors for deliberate use, such as political attitudes or acceptance of traditional gender arrangements. Perceptual effects: GFL can influence perceptions, e.g., using gender-fair language in job descriptions can affect children's perceptions	Characteristics of grammatical gender languages (e.g., German): every noun has grammatical gender; gender of personal nouns typically matches referent; personal pronouns are gendered; pronouns and dependent words signal noun gender; referential gender often explicit but asymmetries exist (e.g., French 'homme' = man/human, German 'alle Wähler' = all voters, masculine for mixed group). GFL policies and implementation: international standards (UNESCO, European Commission) regulate internal documents but are not mandatory. National variations: availability and implementation vary. Mandatory usage: Austria strictly enforces GFL; Germany includes feminine forms in dictionaries and education.
Shrestha, S. and S. Das (2022)	Exploring Gender Biases in ML and AI Academic Research through Systematic Literature Review	Systematic review of gender bias in ML and AI Detailed review of 120 peer-reviewed papers from Google Scholar, ACM, IEEEXplore Filtered for English, accessibility, completeness, relevance to gender bias in automated systems	Key findings in ML/AI: Models reflect societal and data biases Gendered nouns and intersectional biases present in multiple languages Google Translate biased toward male defaults, exceptions for adjectives Occupation words more biased than adjectives Bias observed in AI applications (justice, medical robots, self-driving cars, recommender systems)	The systematic review acknowledges potential limitations, such as missing relevant papers due to technical search constraints (e.g., limited time, keywords, database platforms).	
Smacchia, M., S. Za, and A. Arenas (2024)	Does AI Reflect Human Behaviour? Exploring the Presence of Gender Bias in AI Translation Tools	Investigates gender bias in AI translation tools Focus on languages that allow subject omission translated into languages requiring explicit subjects Objectives: quantify bias, analyse translation method effects, identify jobs prone to bias, compare AI to human behaviour	Two types of bias in AI translation: gender and converging Converging bias: outputs influenced by previous translations Gender-specific bias: male occupations biased toward male forms, female occupations more diverse Tool-specific behaviour: DeepL near-perfect distinction, Google Translate variable, Microsoft Azure biased toward male jobs, GPT-3.5 shows evolving bias over sequences Human responses less biased but still show masculine	Limitations include a small dataset (10 jobs, 30 sentences), limiting generalizability. Future work could expand the dataset, explore more complex sentence structures, other languages (e.g., Spanish, Persian) that allow subject omission, and broader demographic variety in human surveys. Geographical location and iterative request nature could also influence results.	The study confirmed gender bias in AI translation tools reflects underlying training data. Research methodology provides preliminary insights into bias phenomena.
Stanczak, K. and I. Augenstein (2021)	Survey on Gender Bias in Natural Language Processing	Comprehensive survey of 304 papers on gender bias in NLP Summarises developments, identifies limitations, proposes recommendations	◦Recommendations: Diversify metrics, develop standard evaluation benchmarks and tests for comparability, encourage data collection for gender-inclusive task-specific datasets, and address typological variety	Historically, research has been in a strictly binary setting, but there is increasing importance given to gender-inclusive research that accounts for non-binary identities and pronouns.	Nature of gender bias in NLP: stems from implicit sexism in text, biases in model parameters, societal gender gap. NLP models can perpetuate and amplify biases. Most research focuses on English corpora; need for work in other languages with morphological gender agreement. Gender bias increases with model size.
Stanovsky, G., N. A. Smith, and L. Zettlemoyer (2019)	Evaluating Gender Bias in Machine Translation	Uses challenge set with non-stereotypical gender roles to evaluate bias	WinoMT: multilingual automatic evaluation of gender bias 3,888 English sentences from Winogender and WinoBias Evaluates coreference resolution based on roles All tested MT systems showed gender bias Bias follows stereotypes rather than context	The use of gender-swapped adjectives (e.g., "The pretty doctor asked the nurse to help her") to reduce bias was shown to be impractical as a general debiasing scheme, as it assumes oracle coreference resolution.	First foundation: study contributes to work evaluating MT systems using challenge sets, moving beyond BLEU metrics to assess specific linguistic phenomena.
Ullmann, S. (2022)	Gender Bias in Machine Translation Systems	Corpus-linguistic analysis of 17 million English-German sentence pairs Interdisciplinary team (linguistics, philosophy, CS, engineering) analyzed 5% subset to identify gender imbalances Tested techniques to reduce bias in MT system trained on this corpus	German-English bias testing Male pronouns and nouns more frequent than female Pre-training mitigation techniques: downsampling, upsampling, counterfactual augmentation Counterfactual augmentation worked best but still imperfect Conclusion: MT bias persists without intervention, but can be mitigated with little computational effort Interdisciplinary work crucial for long-term solutions		Cause of bias: LLMs trained on vast internet data, often lacking diversity, overrepresenting dominant groups, containing misinformation or harmful language. Systemic bias: structural imbalances transferred via automated processes. Technical bias: system design can constrain data processing, causing unfair distribution. Semantic bias: associative relations (e.g., 'he' + 'doctor'). Amplification of bias occurs during training, exaggerating existing distributions (e.g., cooking associated with women, implying only women cook).