A Appendix

The following sections supplement presented results with further details. Sec. A.1 provides all included gender terms and their frequency. Sec. A.2 presents comprehensive tables with measured biases or all experimental conditions. Sec. A.3 states test accuracies and other evaluation parameters of included classifiers.

A.1 Target Word Sets

Masked Terms The following list presents all gender terms that were, first, removed and masked to create the training conditions *R* and *mix* and, second, masked with an equivalent term of the opposite gender for experimental data. The list was carefully constructed, incorporating previous literature. Bolukbasi et al. (2016) state a comprehensive list of 218 gender-specific words already. We used that as a root and added further terms that we found in the data itself or other sources and that we considered being missing. Our final list comprises 685 terms in total.
In general, if possible, terms were masked by their exact equivalent of the other gender, e.g. *man* by *woman*, and similarly *woman* by *man*. Yet, language and the meaning and connotation of words are highly complex and ambiguous. Thus, the list of terms is not clear-cut, and for some terms, it is disputable whether they should be included or not. These are the four main concerns and how we handled each of them:

First, some mappings are not definite, i.e. there are multiple options to transfer the term into the opposite gender. One example is *lady*, which could be the female version of *gentleman* or *lord*. In these cases, we either selected the most likely translation or randomly.
Second, some terms do not have an appropriate translation like, among others, the term *guy*, or the term does exist in the other gender but is not used (as much), like for the term *feminism*. In these cases, we tried to find any translation that reflects the meaning as accurate as possible, like *gal* for *guy* or applied the rarely used counterpart, e.g. *masculism*.
Third, in some cases, there is a female version of the term, but the male version is usually used for all genders. This is, for example, the case for *manageress* or *lesbianism*. These terms exist and are possibly used, but one could still say 'she is a manager' or 'she is gay'. In these cases, we only translated the term in one direction. This is, whenever the term *lesbian* occurs, it is translated into *gay* for the male version, but when the original rating includes the term *gay*, it is not transformed into *lesbian* for the female version.
Finally, it can have other meanings that are not gender-related, e.g. *Miss* as an appellation can also be the verb *to miss*. We decided to interpret these terms as the more frequent meaning or to leave the term out if it was unclear.
Similar to many other resources, Bolukbasi et al. (2016) also include terms from the animal realm, such as *stud* or *lion*(Bolukbasi et al., 2016). We decided not to do so because the present investigation focuses on human gender bias, which might not be similarly present for animals. The list includes all masked terms that occurred at least ten times in the entire experimental data in decreasing order. Further 404 terms were included in the analysis that occurred fewer than ten times. 221 of these terms were not counted even once and did not affect the analysis. A comprehensive list of all considered terms and their frequency can be found in the corresponding repository.
The full list corresponds to the *all* term set. Due to the above-discussed concerns, we also applied the *weat* term set, which consists of mostly unambiguous terms. Terms that are included in weat are marked in bold. The third term set, *pro*, only includes pronouns which are *he, she, his, her, him* and *hers*. This term set is relatively small, but pronouns are more frequent than most other terms.

Pronouns are marked in bold.

he (46634), **his** (34475), **her** (31303), **she** (26377), **him** (17863), **man** (11656), guys (8070), **girl** (7433), guy (5862), god (5324), mom (4456), actors (4349), **boy** (3802), **girls** (3509), **mother** (3424), dad (3274), **woman** (3235), wife (2858), **brother** (2810), **sister** (2726), **men** (2662), **father** (2468), mr (2439), **boys** (2377), actor (2369), **son** (2226), **women** (2212), himself (2194), dude (2089), **daughter** (1995), lady (1948), husband (1658), boyfriend (1544), **brothers** (1474), hero (1427), actress (1167), **female** (1158), girlfriend (1087), king (1012), **mothers** (1009), hubby (994), count (932), herself (878), **male** (821), daddy (792), ladies (766), ms (725), giant (725), mommy (721), master (708), **sisters** (701), lord (697), ma (671), sir (626), queen (621), mama (596), **uncle** (587), chick (567), moms (556), grandma (529), **aunt**

(521), **fathers** (444), heroes (434), princess (432), pa (411), host (405), niece (373), prince (350), dads (341), actresses (341), priest (328), nephew (328), hunter (303), bride (284), witch (281), lesbian (277), heroine (261), kings (239), grandpa (239), daughters (234), **grandfather** (223), **grandmother** (222), chicks (193), masters (187), cowboy (185), counts (177), dudes (174), sons (169), gods (166), gal (159), papa (158), wifey (156), girly (156), queens (152), bachelor (149), housewives (148), **hers** (148), maid (145), girlfriends (145), beard (141), emperor (136), gentleman (129), superman (128), duke (127), girlie (125), mayor (123), wives (122), gentlemen (116), playboy (114), mister (113), mistress (111), giants (109), females (107), wizard (105), widow (98), nun (98), penis (96), fiance (95), lad (92), gals (92), boyfriends (91), girlies (90), bloke (90), bachelorette (88), aunts (87), policeman (84), males (84), fella (79), diva (79), macho (78), goddess (78), lads (77), landlord (75), fiancé (75), patron (74), waitress (73), husbands (70), hosts (70), fiancée (70), feminist (70), cowboys (70), nephews (68), mermaid (68), sorority (66), grandmas (66), chap (65), manly (64), businessman (63), monk (62), baron (62), witches (61), bachelor (61), nieces (59), housewife (59), **feminine** (58), cameraman (58), shepherd (57), lesbians (55), vagina (53), uncles (53), wizards (52), henchmen (49), salesman (48), postman (48), mamas (48), grandson (48), brotherhood (47), lords (44), henchman (44), waiter (43), dukes (42), mommies (41), fellas (41), granddaughter (40), traitor (39), groom (39), duchess (39), madman (36), policemen (35), conductor (35), sisterhood (34), fraternity (34), monks (33), **masculine** (33), nuns (32), fiancee (32), lass (30), tailor (29), priests (29), maternity (29), butch (29), stepfather (28), hostess (28), ancestors (28), heiress (27), countess (27), congressman (27), bridesmaid (27), protector (26), divas (26), ambassador (26), damsel (25), steward (24), madam (24), homeboy (24), landlady (23), **grandmothers** (23), fireman (23), empress (23), chairman (23), widower (22), sorcerer (22), patrons (22), masculinity (22), firemen (22), englishman (22), businessmen (22), testosterone (21), manhood (21), chaps (21), widows (20), lesbianism (20), blokes (20), beards (20), barbershop (20), anchorman (20), sperm (19), heroines (19), heir (19), stepmother (18), princesses (18), princes (18), handyman (18), patriarch (17), monastery (17), mailman (17), homegirl (17), headmistress (17), fisherman (17), czar (17), brotherly (17), brides (17), uterus (16), maternal (16), abbot (16), prophet (15), boyish (15), adventurer (15), testicles (14), temptress (14), schoolgirl (14), penises (14), maids (14), barmaid (14), waiters (13), traitors (13), stuntman (13), priestess (13), seductress (12), schoolboy (12), motherhood (12), daddies (12), cowgirls (12), cameramen (12), bachelors (12), adventurers (12), sculptor (11), schoolgirls (11), proprietor (11), paternal (11), homeboys (11), foreman (11), feminism (11), doorman (11), bachelors (11), womanhood (10), testicle (10), mistresses (10), merman (10), **grandfathers** (10), girlish (10)

A.2 Biases

Tab. 3 and Tab. 4 provide an overview of the model biases of all considered classifiers. For the calculation of biases, the same gender term set was applied to the experimental data masking as for the training data condition. This means, for instance, in the experimental data for all *R-weat* and *mix-weat* trained classifiers, only *weat* terms were masked. Thus, the training condition is in line with the experimental bias calculation for all *N* and *mix* training conditions. For *original* training conditions, however, no term set was applied to the training data. This is why biases of all three term groups are compared, which are *original-N*, *original-pro*, and *original-weat*.

Wilcoxon signed-rank-test yielded highly significant p-values for almost all conditions. Exceptions are *distbert mix-all*,*bertbase mix-pro*, *robertbase mix-weat*, *robertbase mix-all*, and *albertlarge mix-weat*. Out of 63 reported experimental models, 57 showed highly significant biases, of which 16 prefer female terms over male terms, and 41 prefer male terms over female terms.

Condition	non zero bias abs	non zero bias tot	all bias abs	all bias tot	N< 0	N= 0	N> 0	sign.
distbert								
original-pro	0.0021	0.0009	0.0014	0.0006	6085	10216	8699	***
R-pro	0.0022	0.0010	0.0014	0.0007	7116	9183	8701	***
mix-pro	0.0022	-0.0012	0.0012	-0.0007	6922	7309	10769	***
original-weat	0.0035	0.0004	0.0026	0.0003	8098	10214	6688	***
R-weat	0.0037	-0.0015	0.0027	-0.0011	10773	7532	6695	***
mix-weat	0.0027	-0.0008	0.0018	-0.0006	8332	8428	8240	***
original-all	0.0047	0.0016	0.0039	0.0013	7817	12941	4242	***
R-all	0.0045	0.0008	0.0037	0.0007	10022	10734	4244	***
mix-all	0.0052	-0.0003	0.0042	-0.0003	10080	10177	4743	-
bertbase								
original-pro	0.0025	0.0013	0.0016	0.0008	5430	10874	8696	***
R-pro	0.0036	0.0031	0.0024	0.0020	3234	13061	8705	***
mix-pro	0.0023	-0.0000	0.0014	-0.0000	7505	7794	9701	-
original-weat	0.0037	0.0015	0.0027	0.0011	6187	12128	6685	***
R-weat	0.0038	0.002	0.0028	0.0015	6204	12098	6698	***
mix-weat	0.0027	-0.0002	0.0015	-0.0001	6421	7135	11444	***
original-all	0.0056	0.0035	0.0046	0.0029	5233	15527	4240	***
R-all	0.0060	0.0041	0.0049	0.0034	4319	16431	4250	***
mix-all	0.0055	0.0005	0.0035	0.0003	6838	9001	9161	***
bertlarge								
original-pro	0.0031	-0.0016	0.0021	-0.0011	10287	6020	8693	***
R-pro	0.0050	0.0046	0.0032	0.0030	2697	13610	8693	***
mix-pro	0.0035	0.0011	0.0014	0.0004	4961	5228	14811	*
original-weat	0.0069	-0.0032	0.0051	-0.0023	10329	7986	6685	***
R-weat	0.0048	-0.0011	0.0035	-0.0008	10172	8142	6686	***
mix-weat	0.0056	0.0034	0.0029	0.0018	4581	8195	12224	***
original-all	0.0082	0.0009	0.0068	0.0007	9128	11633	4239	***
R-all	0.0095	0.0042	0.0079	0.0035	7314	13443	4243	***
mix-all	0.0101	0.0015	0.0078	0.0012	8848	10455	5697	***

Table 3: Total biases of all experimental classifiers (part 1). The bias is the mean bias over all experimental samples. While the absolute bias (bias abs) is the mean of absolute values, the total bias (bias tot) is based on the directed sample biases. For “non zero" values, samples with a bias$= 0$ are excluded. “all" includes all 25000 sample biases. The numbers of samples with negative, no, and positive bias are given by $N < 0$, $N = 0$, or $N > 0$, respectively. Significance levels for Wilcoxon signed-rank-test were defined as $p > 0.05$:*, $p > 0.01$:**, and $p > 0.001$:***. Reported significance levels were corrected for multiple testing with the Bonferroni correction.

	non zero		all					
Condition	bias abs	bias tot	bias abs	bias tot	N< 0	N= 0	N> 0	sign.
robertabase								
original-pro	0.0024	0.0016	0.0015	0.0010	5448	10840	8712	***
R-pro	0.0024	0.0009	0.0015	0.0006	6822	9472	8706	***
mix-pro	0.0021	-0.0002	0.0013	-0.0001	8682	7612	8706	***
original-weat	0.0031	0.0016	0.0023	0.0011	6470	11832	6698	***
R-weat	0.0028	0.0007	0.0021	0.0005	7722	10581	6697	***
mix-weat	0.0023	0.0002	0.0017	0.0002	9396	8894	6710	-
original-all	0.0036	0.0020	0.0030	0.0016	7165	13585	4250	***
R-all	0.0038	0.0010	0.0032	0.0008	9294	11464	4242	***
mix-all	0.0027	0.0000	0.0023	0.0000	10520	10206	4274	-
robertalarge								
original-pro	0.0024	0.0015	0.0016	0.0010	5235	11055	8710	***
R-pro	0.0025	0.0015	0.0016	0.0010	5216	11072	8712	***
mix-pro	0.0020	0.0004	0.0013	0.0003	6679	9606	8715	***
original-weat	0.0039	0.0025	0.0029	0.0018	5894	12411	6695	***
R-weat	0.0039	0.0023	0.0029	0.0017	6109	12193	6698	***
mix-weat	0.0028	0.0004	0.0021	0.0003	8071	10220	6709	***
original-all	0.0044	0.0023	0.0036	0.0019	7105	13653	4242	***
R-all	0.0043	0.0021	0.0035	0.0017	7045	13712	4243	***
mix-all	0.0041	0.0018	0.0034	0.0015	6971	13783	4246	***
albertbase								
original-pro	0.0037	0.0011	0.0024	0.0007	5481	10811	8708	***
R-pro	0.0029	-0.0004	0.0019	-0.0003	9305	6986	8709	***
mix-pro	0.0054	0.0021	0.0035	0.0014	7244	8968	8788	***
original-weat	0.0093	0.0002	0.0068	0.0001	7710	10600	6690	***
R-weat	0.0082	-0.0044	0.006	-0.0032	10346	7942	6712	***
mix-weat	0.0131	-0.0034	0.0093	-0.0024	9426	8263	7311	***
original-all	0.0089	-0.0023	0.0074	-0.0019	9112	11645	4243	***
R-all	0.0080	0.0009	0.0067	0.0008	8979	11769	4252	***
mix-all	0.0071	-0.0014	0.0058	-0.0012	9481	11030	4489	***
albertlarge								
original-pro	0.0086	0.0086	0.0056	0.0056	2120	14075	8805	***
R-pro	0.0049	0.0034	0.0032	0.0022	6407	9869	8724	***
mix-pro	0.0016	-0.0008	0.0010	-0.0005	10121	6136	8743	***
original-weat	0.0155	0.0130	0.0113	0.0095	4058	14191	6751	***
R-weat	0.0074	-0.0032	0.0054	-0.0023	9936	8373	6691	***
mix-weat	0.0091	-0.0009	0.0066	-0.0006	9186	8998	6816	-
original-all	0.0172	0.0137	0.0143	0.0114	5095	15594	4311	***
R-all	0.0114	-0.0032	0.0095	-0.0026	12573	8180	4247	***
mix-all	0.0101	0.0034	0.0084	0.0028	8777	11875	4348	***

Table 4: Total biases of all experimental classifiers (part 2). Extension of Tab. 3

Model / Spec	acc.	rec.	prec.	f1
distbase				
original	.812	.778	.835	.805
R-all	.817	.789	.836	.812
R-weat	.820	.789	.840	.814
R-pro	.818	.780	.844	.811
mix-all	.822	.795	.840	.817
mix-weat	.822	.783	.849	.815
mix-pro	.822	.784	.848	.815
bertbase				
original	.818	.787	.838	.812
R-all	.821	.781	.849	.813
R-pro	.820	.776	.851	.812
R-weat	.821	.803	.833	.818
mix-all	.836	.791	.868	.828
mix-pro	.835	.816	.849	.832
mix-weat	.835	.812	.852	.832
bertlarge				
original	.805	.787	.816	.801
R-all	.797	.734	.839	.783
R-pro	.779	.660	.867	.749
R-weat	.803	.739	.847	.789
mix-all	.795	.723	.845	.780
mix-pro	.797	.738	.836	.784
mix-weat	.789	.710	.843	.771

Table 5: Test accuracy (acc.), recall (rec.), precision (prec.), and F1-Score (f1) for the models that are used in the experiments - part 1

A.3 Evaluation of Models

Tab. 5 and Tab. 6 show the accuracies, recalls, precisions and F1-Score of all experimental models calculated on the test data. For the calculation of reported values, the test data set has been treated analogously to the training condition. That means for instance, since we removed all pronouns from training data in the R-all condition, we did the same in the test data before evaluating the models in that condition.

Model / Spec	acc.	rec.	prec.	f1
robertabase				
original	.818	.744	.874	.804
R-all	.823	.770	.862	.813
R-weat	.820	.739	.881	.804
R-pro	.818	.733	.883	.801
mix-all	.833	.780	.873	.824
mix-weat	.830	.781	.867	.821
mix-pro	.823	.760	.870	.811
robertalarge				
original	.820	.748	.873	.806
R-all	.820	.765	.859	.810
R-weat	.820	.761	.862	.809
R-pro	.818	.751	.868	.805
mix-all	.815	.749	.862	.801
mix-weat	.816	.761	.855	.805
mix-pro	.814	.728	.879	.797
albertbase				
original	.693	.932	.630	.752
R-all	.771	.711	.809	.756
R-weat	.772	.749	.785	.767
R-pro	.757	.748	.764	.756
mix-all	.782	.791	.777	.784
mix-weat	.778	.818	.757	.786
mix-pro	.780	.813	.762	.787
albertlarge				
original	.784	.762	.797	.779
R-all	.762	.847	.724	.781
R-weat	.767	.802	.750	.775
R-pro	.763	.832	.732	.779
mix-all	.774	.803	.759	.781
mix-weat	.784	.788	.781	.785
mix-pro	.782	.752	.801	.776

Table 6: Test accuracy (acc.), recall (rec.), precision (prec.), and F1-Score (f1) for the models that are used in the experiments - part 2