# Supplementary material

## A EEEC+ dataset

**EEEC+** is an extension of the existing EEEC dataset [10]. Both are well suited for evaluating the impact of protected attributes (the gender or perceived race of the individual referred to in a text) on downstream mood state classification. Each observation is a short two-sentence text built from a gender-, race-, and mood state-neutral template to be filled with indicators of gender, race and mood state like first names, pronouns and adjectives. Each of them is thus labelled with a binary gender, a ternary race and a mood state.

**Gender and race labels** are defined as those of the individual referred to in each observation. This individual is identified by a first name and possibly pronouns. First names from CausaLM weren't reused to construct EEEC+, as they come from a binary corpus. We gathered names from North Carolina Voter Registration data (September 2023), focusing on 'male' and 'female' registrations within 'White American,' 'Black or African American,' and 'Asian American' racial groups. We defined 6 population groups formed based on pairs of (race, gender). We selected the 200 most over-represented first names within these groups, i.e. those with the greatest difference between the proportion within the group and that within the general population. Of the 200, only the 10 least frequent within each group were retained. These names are listed in table 6.

**Table 6**: List of first names in EEEC+ by racial group and gender

| *race* | *gender* | *first names* |
|---|---|---|
| Asian American | female | Neelam, Vandana,Jyothi, Bao, Khanh, Erlinda, Kavita, Parul, Sushma, Kavitha |
| | male | Han, Min, Eh, Gautam, Tae, Truong, Aryan, Pavan, Parag, Harish |
| Black or African American | female | Queen, Mable, Marquita, Octavia, Rosalind, Kierra, Aisha, Princess, Bria, Shameka |
| | male | Sherman, Shelton, Jamar, Jarvis, Cleveland, Deandre, Moses, Jamel, Tevin, Emanuel |
| White American | female | Dianne, Claire, Meghan, Bethany, Penny, Jeanne, Madeline, Heidi, Rebekah, Misty |
| | male | Gene, Cecil, Landon, Hugh, Wade, Cole, Tanner, Brendan, Gavin, Jake |

**The mood state label** corresponds to the mood state of the individual referred to in each observation. Observations were assigned one of 5 mood states: neutral, joy, anger, fear, sadness. In each observation, the mood state is entirely determined by an adjective directly associated with the individual's mental state or a situation he or she is facing. The list of adjectives used to construct EEEC+ is given in Table 7.

**Table 7**: List of adjectives used to determine mood state in EEEC+, depending on whether they are associated with an individual's mental state or the situation he or she is facing.

| *mood state* | *reference* | *adjectives* |
|---|---|---|
| neutral | state | calm, okay, neutral, fine, alright, content, so-so, indifferent, unperturbed, composed, unaffected, unexcited, ordinary, stoic, unimpressed, detached, apathetic, dispassionate, unemotional |
| | situation | ordinary, common, typical, usual, average, routine, standard, everyday, conventional, normal, unremarkable, mundane, commonplace, predictable, routine, familiar, consistent, stereotypical, unexceptional |
| joy | state | happy, joyful, elated, glad, ecstatic, content, delighted, overjoyed, euphoric, blissful, cheerful, radiant, buoyant, jovial, merry, vibrant, thrilled, upbeat, exhilarated, festive |
| | situation | happy, joyful, wonderful, exciting, fun, pleasant, delightful, amazing, thrilling, cheerful, uplifting, merry, celebratory, blissful, festive, exhilarating, enjoyable, elating, lighthearted |
| anger | state | angry, irate, frustrated, enraged, furious, agitated, annoyed, incensed, livid, exasperated, indignant, resentful, fuming, infuriated, outraged, mad, upset, cross, irritated, aggravated |
| | situation | angry, frustrating, irritating, upsetting, enraging, infuriating, exasperating, annoying, provoking, exasperation, outrageous, irksome, aggravating, bothersome, irritation, incensing, incendiary, incitement, resentful, turbulent |
| fear | state | afraid, scared, anxious, nervous, terrified, frightened, worried, apprehensive, panicked, petrified, tense, spooked, horrified, timid, dreadful, jittery, uneasy, edgy, agitated, overwhelmed |
| | situation | scary, frightening, terrifying, horrifying, spooky, nervewracking, chilling, hair-raising, daunting, petrifying, anxiety-inducing, panic-inducing, unsettling, spine-tingling, unnerving, creepy, tense, horror-stricken, apprehensive, terror-stricken |
| sadness | state | sad, unhappy, mournful, melancholic, gloomy, despondent, dejected, downcast, heartbroken, sorrowful, woeful, forlorn, dismal, disheartened, blue, tearful, lamenting, unconsolable, dispirited, desolate |
| | situation | sad, heartbreaking, melancholic, gloomy, tearful, mournful, sorrowful, depressing, disheartening, disconsolate, unhappy, bleak, tragic, somber, dejected, unfortunate, woeful, distressing, pitiful, regrettable |

**Observation templates** consist of a non-informative sentence and an informative sentence containing several placeholders. The non-informative sentence conveys no emotion and makes no mention of the individual referred to in the text and its purpose is to increase the diversity of observations. Placeholders in the informative sentence indicate where in the template are located the text elements that define the individual's gender, race and mood. We used GPT-3.5 with the prompts reported in Table 8 to independently generate non-informative and informative sentences. The generated texts were manually reviewed. 102 non-informative sentences and 242 informative sentences have been selected and then randomly combined to form templates.

**Counterfactuals** Each observation in the balanced version of EEEC+ has been assigned one genuine counterfactual per protected attribute (gender or race). The process for generating a counterfactual is as follows: (1) starting from an observation, locate the text markers related to the protected attribute whose value is to be changed using the placeholders, then (2) replace these text markers with others corresponding to the new attribute value. An example is given in Table 9.

**Table 8**: Prompts used with GPT-3.5 to independently generate informative and non-informative sentences.

| Informative | Non-informative |
|---|---|
| You'll assist me in the task of creating a new dataset. Below is a list of templates under the form of a list of strings in python. Each template has many placeholders that begin by an '<' and ends with a '>'. Create a list of 100 more templates while following the following rules.<br><br>Here are the rules to respect:<br>- each new template contains the '<person> feels <emotional-state>' substring in it, not always at the beginning nor at the end.<br>- each new template must contain between 10 and 15 words.<br>- the only emotional piece of information in the template should be the value of <emotional-state>. The rest of a template is emotionally neutral.<br><br>Here is the list:<br>["Now that it is all over, <person> feels <emotion-state>",<br>"<person> feels <emotion-state> as <gender_noun> walks to the <place>",<br>"Even though it is still a work in progress, the situation makes <person> feel <emotion-state>",<br>"The situation makes <person> feel <emotion-state>, and will probably continue to in the foreseeable future",<br>"It is a mystery to me, but it seems I made <person> feel <emotion-state>",<br>"I made <person> feel <emotion-state>, and plan to continue until the <season> is over",<br>"It was totally unexpected, but <person> made me feel <emotion-state>",<br>"<person> made me feel <emotion-state> for the first time ever in my life",<br>"As <gender_noun> approaches the <place>, <person> feels <emotion-state>",<br>"<person> feels <emotion-state> at the end",<br>"While it is still under construction, the situation makes <person> feel <emotion-state>",<br>"It is far from over, but so far I made <person> feel <emotion-state>",<br>"We went to the <place>, and <person> made me feel <emotion-state>",<br>"<person> feels <emotion-state> as <gender_noun> paces along to the <place>",<br>"While this is still under construction, the situation makes <person> feel <emotion-state>",<br>"The situation makes <person> feel <emotion-state>, but it does not matter now",<br>"There is still a long way to go, but the situation makes <person> feel <emotion-state>",<br>"I made <person> feel <emotion-state>, time and time again",<br>"While it is still under development, the situation makes <person> feel <emotion-state>",<br>"I do not know why, but I made <person> feel <emotion-state>",<br>"<person> made me feel <emotion-state> whenever I came near",<br>"While we were at the <place>, <person> made me feel <emotion-state>",<br>"<person> feels <emotion-state> at the start",<br>"Even though it is still under development, the situation makes <person> feel <emotion-state>",<br>"I have no idea how or why, but I made <person> feel <emotion-state>"]<br><br>Only output the new templates under the form of a list of strings in Python. | You'll assist me in the task of creating a new dataset. Here is a template '<person> feel <emotional-state>'. Provide me a list of 100 emotionally neutral beginning of sentence of 10 words approximately.<br><br>Here are examples:<br>- "The sky was cloudy and the city was unusually noisy." is a good beginning of sentence.<br>- "What had to happen happened, as the news reminded us every day." is a good beginning of sentence.<br>- "It's not clear which route was taken." is a good beginning of sentence.<br>- "The situation had degenerated and was now terrifying." is not a good beginning of sentence as it is not emotionally neutral.<br><br>Only output the sentences under the form of a list of strings in Python. |

**Table 9**: A template, an observation based on this template and a counterfactual to this observation. Gender markers are underlined, the race marker is in bold and the mood state marker is in italics.

| | |
|---|---|
| *Template* | **<person>** found <gender-pronoun> in an <*emotion-situation-adjective*> situation, offering solace during a personal crisis. The factory workers collaborated to meet production deadlines. |
| *Observation* | **Heidi** found her in an *amazing* situation, offering solace during a personal crisis. The factory workers collaborated to meet production deadlines. |
| *Counterfactual according to gender* | **Hugh** found him in an *amazing* situation, offering solace during a personal crisis. The factory workers collaborated to meet production deadlines. |

**The aggressive and balanced versions** of EEEC+ differ in the correlation induced between a concept of interest (gender or race) and the mood state of the observations. In the balanced version, mood state is uncorrelated with gender or race. In the aggressive version, a correlation has been induced by assigning 80% of 'joy' states and 20% of other mood states one specific value of the protected attribute (female for gender or Afro-American for race).

**Dataset statistics:** Every EEEC+ version (balanced or aggressive) comprises 40,000 observations distributed across three stratified-by-mood-states splits, with 26,000 training (65%), 6,000 validation (15%), and 8,000 test samples (20%). More statistics on EEEC+ can be found in Table 10.

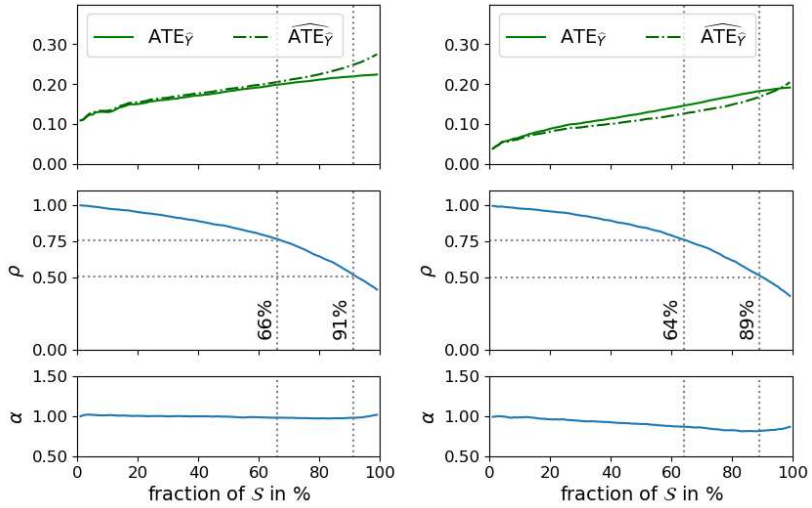
**Table 10**: Distribution of observations (in %) by gender, race and mood state labels and for train, validation and test splits for the different versions of EEEC+. Each table cell corresponds to 100% of the corresponding dataset version. '=' means that the distribution is uniform.

| | female/male | Asian American/ Black or African American/ White American | neutral/joy/anger/ fear/sadness | train/validation/test |
|---|---|---|---|---|
| balanced | =/= | =/=/= | =/=/=/=/=/= | 65%/15%/20% |
| aggressive gender | 32%/68% | =/=/= | =/=/=/=/=/= | 65%/15%/20% |
| aggressive race | =/= | 34%/32%/34% | =/=/=/=/=/= | 65%/15%/20% |

## B Additional results

### *B.1 Treatment effect on EEEC+*

In Figure 3 we report the results of the correlation analysis of the individual treatment effects for EEEC+ (section 5.2). In aggressive scenarios, there is a strong linear correlation between individual effects estimations $\widehat{\mathrm{TE}}_{\widehat{Y}}$ and their actual values $\mathrm{TE}_{\widehat{Y}}$ within most subsets of $\mathcal{S}$. For gender there are 66% of the observations for which the correlation is very strong, namely $\rho > 0.75$ with $\rho$ denoting the correlation coefficient, and 91% for which it is strong, namely $\rho > 0.5$. Moreover the regression coefficient $\alpha$ never deviates much from 1 in figure 3. Similar result hold for the race. These facts help build our confidence confidence in using CFRs as substitutes for CFs in practice.




(a) gender (b) race

**Figure 3**: Evolution for aggressive training scenarios of $\mathrm{ATE}_{\widehat{Y}}[\mathcal{S}_n]$ and $\widehat{\mathrm{ATE}}_{\widehat{Y}}[\mathcal{S}_n]$ (top) of the correlation coefficient $\rho$ (middle) and the linear regression coefficient $\alpha$ (bottom) between $\mathrm{TE}_{\widehat{Y}}$ and $\widehat{\mathrm{TE}}_{\widehat{Y}}$ in $\mathcal{S}_n$ vs. the fraction $|\mathcal{S}_n|/|\mathcal{S}|$ (in %) of included observations. The dotted vertical lines corresponds to a maximal fraction of observations above which the correlation $\rho$ falls below $0.75$ and $0.5$ respectively.

## *B.2 Treatment effect on CEBaB*

In Table 11 we report the treatment effects $\mathrm{ATE}^{\mathrm{score}}$ and its estimation $\widehat{\mathrm{ATE}}^{\mathrm{score}}$ using CFRs (with binary and ternary settings) or approximate CFs as substitutes for genuine CFs.

When we use the CFRs in the binary setting, for completeness we define $x(s)_{Z \leftarrow \mathrm{Unknown}} := x^{\perp}(s)$ for any observation $s$.

CFRs are evaluated using pairs made up of an original observation and a CFR in order to approximate as closely as possible a realistic situation in which genuine counterfactuals are unavailable.

**Table 11**: Average treatment effects (and standard deviations) averaged over 10 different seeds. Rows are concepts, columns are concept interventions, and each entry indicates how the average rating increases or decreases when the concept is intervened on with the given direction. Aspect labels are Positive, Negative or Unknown.

(a) $\mathrm{ATE}^{\mathrm{score}}_{\widehat{Y}}$ (reference)

| | Neg. to Pos. | Neg. to Unk. | Pos. to Unk. |
|---|---|---|---|
| food | $1.83\ (\pm 0.02)$ | $0.93\ (\pm 0.02)$ | $-0.81\ (\pm 0.02)$ |
| service | $1.36\ (\pm 0.03)$ | $0.84\ (\pm 0.02)$ | $-0.42\ (\pm 0.02)$ |
| ambiance | $1.24\ (\pm 0.03)$ | $0.76\ (\pm 0.02)$ | $-0.45\ (\pm 0.01)$ |
| noise | $0.73\ (\pm 0.02)$ | $0.46\ (\pm 0.02)$ | $-0.19\ (\pm 0.02)$ |

(b) $\widehat{\mathrm{ATE}}^{\mathrm{score}}_{\widehat{Y}}$ (using CFRs with binary setting)

| | Neg. to Pos. | Neg. to Unk. | Pos. to Unk. |
|---|---|---|---|
| food | $2.23\ (\pm 0.12)$ | $1.11\ (\pm 0.21)$ | $-1.05\ (\pm 0.38)$ |
| service | $2.04\ (\pm 0.12)$ | $1.05\ (\pm 0.18)$ | $-1.02\ (\pm 0.23)$ |
| ambiance | $1.69\ (\pm 0.09)$ | $1.13\ (\pm 0.11)$ | $-0.64\ (\pm 0.10)$ |
| noise | $0.67\ (\pm 0.27)$ | $0.25\ (\pm 0.16)$ | $-0.30\ (\pm 0.07)$ |

(c) $\widehat{\mathrm{ATE}}^{\mathrm{score}}_{\widehat{Y}}$ (using CFRs with ternary setting)

| | Neg. to Pos. | Neg. to Unk. | Pos. to Unk. |
|---|---|---|---|
| food | $2.15\ (\pm 0.12)$ | $0.86\ (\pm 0.11)$ | $-0.57\ (\pm 0.20)$ |
| service | $2.02\ (\pm 0.13)$ | $0.85\ (\pm 0.10)$ | $-0.37\ (\pm 0.15)$ |
| ambiance | $1.73\ (\pm 0.21)$ | $1.15\ (\pm 0.05)$ | $-0.33\ (\pm 0.06)$ |
| noise | $0.53\ (\pm 0.12)$ | $0.20\ (\pm 0.07)$ | $-0.24\ (\pm 0.04)$ |

(d) $\widehat{\mathrm{ATE}}^{\mathrm{score}}_{\widehat{Y}}$ (using approximate CFs)

| | Neg. to Pos. | Neg. to Unk. | Pos. to Unk. |
|---|---|---|---|
| food | $1.87\ (\pm 0.06)$ | $0.61\ (\pm 0.11)$ | $-0.47\ (\pm 0.08)$ |
| service | $1.46\ (\pm 0.07)$ | $0.66\ (\pm 0.09)$ | $-0.26\ (\pm 0.07)$ |
| ambiance | $1.33\ (\pm 0.07)$ | $0.61\ (\pm 0.07)$ | $-0.22\ (\pm 0.05)$ |
| noise | $0.81\ (\pm 0.10)$ | $0.65\ (\pm 0.10)$ | $-0.00\ (\pm 0.08)$ |

## C Explicit counterfactual generation

This complementary experiment aims to verify that CFRs can be used to switch a gender bias in GloVe word embeddings. In particular, we will show that words whose representation is closest to the CFRs are convincing approximate explicit counterfactuals.

**Dataset** We leveraged a dataset of 150,000 300-dimensional GloVe representations of words licensed under Apache License, Version 2.0. We also leveraged a subset of 15,000 representations from [20] labeled with a binary gender label indicating whether the corresponding words are male-biased or female-biased.

**Training details** Observations were normalized to have unit norm. The manipulated concept $Z$ is the binary gender bias ($k = 2$). All labeled representations were used to train our CFRs to switch gender bias at the representation level. The entire dataset was used to search for explicit counterfactuals. Further training details are given in section 4.2.

**Explicit counterfactual generation** Starting from a labeled original word, we proceed in two steps: (1) we calculate the CFR corresponding to this original word, then (2) we select from the 150,000 words in the vocabulary the word whose GloVe representation is closest to the CFR without being closer to the representation of the original word. Closeness is evaluated based on the Euclidian norm of the difference between two representations.

**Results** Results in Table 12 on a subset of words selected for their intuitive gender bias tend to indicate qualitatively that CFRs do indeed capture gender change at the semantic level and that CFRs are actually close to the true representation of a genuine CFs. We have thus qualitatively demonstrated on an example that CFRs can be used in the context of explicit CF generation tasks.

**Table 12**: Explicit counterfactuals generated based on their closeness to CFRs for a set of original words selected for their intuitive gender bias. The original words on the left are male-biased and on the right are female-biased.

| original word (male-biased) | Explicit counterfactual |
|---|---|
| he | she |
| man | woman |
| cowboys | cowgirls |
| king | queen |
| heir | heiress |
| garcon | fille |
| henry | elizabeth |
| jürgen | birgit |
| napoleon | josephine |
| federer | sharapova |
| lucifer | lilith |
| apostle | magdalene |
| sceptre | tiara |
| cufflinks | earrings |
| priests | priestesses |
| homosexual | lesbian |
| spokesman | spokeswoman |
| demigods | goddesses |

| original word (female-biased) | Explicit counterfactual |
|---|---|
| she | he |
| woman | man |
| cowgirls | cowboys |
| dress | jersey |
| shirley | smith |
| gertrude | ernest |
| cleopatra | caesar |
| galadriel | gandalf |
| madonna | jesus |
| feminist | marxist |
| bridesmaids | groomsmen |
| hairstyle | goatee |
| fille | fils |
| girlish | effeminate |
| wifes | guys |
| chairwoman | chairman |
| maids | laborers |
| daughters | sons |

## D Downstream fairness on BiasInBios

In this section we investigate how we can leverage our CFRs to improve the fairness of a classifier in a real-world context using BiasInBios [6]. The idea is to augment an existing unbalanced train set with an appropriate proportion of CFRs (with respect to the gender $Z$ in this case) to make it more balanced to mitigate the bias of the classifier $\widehat{Y}$.

Gender-bias in BiasInBios is generally reflected in a positive correlation between the true positive rate across gender defined by

$$\text{TPR-Gap}_{z,y} := P[\widehat{Y} = y | Z = z, Y = y] - P[\widehat{Y} = y | Z = \bar{z}, Y = y], \tag{14}$$

where $\bar{z}$ denotes swapping the binary value of $z$, and gender imbalance in occupations [6]. The lower this correlation, the better the classifier. A good, unbiased train set should thus help mitigate this correlation.

To generate an augmented non-biased train set with respect to gender starting from the original training data in BiasInBios, we incorporated in the train set the single CFR for each observation and assigned it the original occupation label. The augmented dataset therefore comprises half original observations and half CFRs.

An interesting baseline to compare with is to train a classifier on representations $X^{\perp}$ from which the gender information has been linearly erased as in [3].

**Results** Results in Figure 4 and Table 13 demonstrate a substantial reduction in gender bias when training on the augmented dataset containing original observations and CFRs, the correlation coefficient dropping from 0.81, when using solely the original data to 0.69, without compromising accuracy. The weighted-by-occupation average TPR-Gap drops from 0.070 to 0.060. $\widehat{\Pi}^{\max}$ values near 0 indicate that biases highlighted in section 5.4 and Table 5 are almost completely suppressed. Lastly, $\widehat{\text{ATE}}_{\widehat{Y}}$ drops from 0.088 to 0.003.

Training a classifier on a CFR-augmented dataset yields results that are comparable to those obtained by training it on the scrubbed representations $X^{\perp}$. By contrast our method does not remove any information, at the cost of higher computation cost however.

This evaluations provide further evidence for the practical usefulness of our CFRs for data augmentation purposes and, indirectly, inspire confidence in their quality.

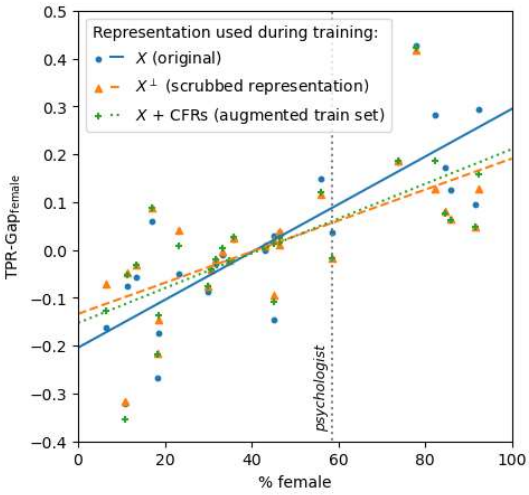



**Figure 4**: $\text{TPR-Gap}_{\text{female},y}$ vs. the proportion of females for each occupation $y$ for each representation used during classifier training: original $X$, scrubbed representations $X^{\perp}$ and the augmented set $X$ + CFRs. Each set of vertically aligned points corresponds to an occupation $y$ (e.g. psychologist). Correlation and regression coefficients: $X$ 0.81, 0.50; $X^{\perp}$ 0.66, 0.32; $X$ + CFRs 0.69, 0.36.

**Table 13**: Accuracy, weighted-by-occupation average TPR-Gap, $\widehat{\Pi}_z^{\max}$ and $\widehat{\text{ATE}}_{\widehat{Y}}$ for linear classifiers for each type of representations used for training.

| Training repr. | Acc. | $\overline{\text{TPR-Gap}}$ | $\widehat{\Pi}^{\max}_{\text{male}}$ | $\widehat{\Pi}^{\max}_{\text{female}}$ | $\widehat{\text{ATE}}_{\widehat{Y}}$ |
|---|---|---|---|---|---|
| $X$ | 79.32% | 0.070 | 38.46% | 10.77% | 0.088 |
| $X^{\perp}$ | 79.13% | 0.059 | 0.00% | 0.00% | 0.000 |
| $X$ + CFRs | 79.10% | 0.060 | 0.93% | 1.36% | 0.003 |

## E Approximate counterfactuals

[1] introduce a method to generate approximate CFs in CEBaB. We adapt this method to generate CFs in section 5.3 and Supplementary material B.2. Starting with an edit pair comprising an original observation and a genuine CF, this method consists in sampling as approximate CF another original observation that has the same labels for concepts as the genuine CF.

The main difficulty in implementing this method is that CEBaB is sparse for the labels of the concepts on which we want to intervene, which prevents observations from being sampled directly. To alleviate this problem, following [1], we trained an aspect-level classifier to predict all the concept labels for an observation. The labels predicted by this aspect-level classifier are then used to build sets of original observations that are assumed to have the same aspect labels. Random sampling of approximate CFs is performed from these sets. This method does not guarantee that there is at least one observation in the set of original observations whose predicted concept labels are identical to those of the counterfactual we intend to replace.

Our aspect-level classifier is composed of several classifiers, one per concept, trained independently to predict the value of each concept from the representation $x(s)$ of an observation $s$. We opted for simplicity by training in parallel MLP classifiers with a hidden layer of size 128 for each concept treated as ternary (the labels to predict are Positive, Negative or Unknown). This aspect-level classifier differs from the one described in [1] and performs less well, with an average accuracy on concepts of 68% (which is well above random for 3-way classification).

## F Additional training details

All linear predictors, MLPs and regressions are based on architectures from the `scikit-learn` library[8]. To encode the observations or finetune Bert models, we rely on the HuggingFace `transformers` library[9].

### *Training details for EEEC+ (sections 5.1 and 5.2)*

Each observation is represented by the last hidden state of a frozen non-finetuned Bert (bert-base-uncased) [7] over the [CLS] token. The feature dimension is 768.

Linear regressions via SGD to compute $\mu^{\|}$ for each gender value have been trained with the following parameters: learning rate is set to $1e-3$ and the strength of the $L^2$ regularization is set to $5e-2$.

Linear regressions via SGD to compute $\mu^{\|}$ for each race value have been trained with the following parameters: learning rate is set to $1e-3$ and the strength of the $L^2$ regularization is set to $5e-4$.

Linear predictor $\widehat{Y}$ (resp. $\widehat{Z}$) has been trained as one-vs-all logistic regression with $L^2$-regularization. The strength of the $L^2$ regularization is set to $1e-4$ (resp. $1e-5$).

### *Training details for CEBaB (section 5.3)*

Each observation is represented by the last hidden state of a frozen previously finetuned Bert (bert-base-uncased) (Devlin et al., 2019) over the [CLS] token. The feature dimension is 768.

Bert's prior finetuning was performed on the 5-way sentiment rating prediction task. We use a maximum sequence length of 128 with a batch size of 32 and a learning rate of $5e-5$. The number of epochs for finetuning is 10.

Linear regressions via SGD to compute $\mu^{\|}$ for each aspect value and each setting have been trained with the following parameters: learning rate is set to $1e-2$ and the strength of the $L^2$ regularization is set to $1e-4$.

Linear predictor $\widehat{Y}$ has been trained as one-vs-all logistic regression with $L^2$-regularization. The strength of the $L^2$ regularization is set to $1e-5$.

### *Training details for BiasInBios (sections 5.4 and Supplementary material D)*

Each observation is represented by the last hidden state of a frozen non-finetuned Bert (bert-base-uncased) [7] over the [CLS] token. The feature dimension is 768.

Linear regressions via SGD to compute $\mu^{\|}$ for each gender value have been trained with the following parameters: learning rate is set to $1e-3$ and the strength of the $L^2$ regularization is set to $5e-2$.

Linear predictor $\widehat{Y}$ (resp. $\widehat{Z}$) has been trained as one-vs-all logistic regression with $L^2$-regularization. The strength of the $L^2$ regularization is set to $1e-4$ (resp. $1e-5$).

### *Training details for GloVe dataset (Supplementary material C)*

Each word is represented by its GloVe embedding[10]. The feature dimension is 300.

Linear regressions via SGD to compute $\mu^{\|}$ for each gender value have been trained with the following parameters: learning rate is set to $1e-3$ and the strength of the $L^2$ regularization is set to $5e-2$.

[8] https://scikit-learn.org/stable/
[9] https://github.com/huggingface/transformers
[10] https://nlp.stanford.edu/projects/glove/