Appendix A. Race and Ignorability

The analysis in this paper using selection did not include race and this is because we do not feel it correct to make the assumption of ignorability at any stage when race is our sensitive attribute. Therefore Assumption 1 would be misguided.

This is because race is correlated with many covariates and it is unclear in general what we are trying to counterfactually correct for. This point relates more to recent philosophical work on counterfactual fairness (Kasirzadeh and Smart, 2021; Hu and Kohler-Hausmann, 2020; Kohler-Hausmann, 2018) as well as work on race in causal studies (Sen and Wasow, 2016). This is beyond the scope of this paper as we have raised challenges for counterfactual fairness methods from within the causal framework, as opposed to the work referenced which raises problems with the use of the causal framework in this setting. However we give a brief example to highlight why we did not include race.

In America and many other Western nations, the race a child is born to is correlated with many crucial demographic features; these include the level of education in their family, socioeconomic status and the neighbourhood of their birth. This makes any comparisons to randomised control trials seem far fetched as these features are likely to affect almost all outcomes we measure in later life. Should our counterfactuals correct for this or not? This relates to what philosophical definition one uses of race, the point made in 'Race as a bundle of sticks' (Sen and Wasow, 2016). The perspective of a racial constructivist, the most popular view in the social sciences, says that racial categories are not a biological fact but they are a social reality and they are inextricably tied with historic 'differences in resources, opportunities, and well-being' (Zalta et al., 1995). Therefore maybe our counterfactuals should correct for this. However taking this point of view it is not clear how we should interpret any counterfactual or if the causal framework fairness for race makes sense as Kohler-Hausmann (2018); Hu and Kohler-Hausmann (2020) and Kasirzadeh and Smart (2021) point out.

If one instead takes an essentialist point of view (this is largely unpopular in the social sciences but it is often implicitly assumed in causal studies) then potentially not, but then we clearly will not satisfy ignorability as we have features that are not caused by race at birth but are correlated with it. (Kusner et al., 2017) mention possibly including as parents' race as a causal ancestor of race in our DAG and also having this as a protected attribute. However, we again run into the same problems as parents' race will be correlated with the same features but one generation back. Therefore when, if ever, would we be happy to say our data could be described by a causal graph with ancestrally closed sensitive attribute set and independent noise variables? This tracking of features back generations in an effort to counterfactually correct for them once again relates more to racial constructivist perspectives, since we are struggling to separate race as something to correct for the discriminatory effects for from the historic context that created it.

We also note these perspectives also apply to gender and other sensitive attributes. However as we mention at the start of the appendix we have raised challenges for counterfactual fairness methods from within the causal framework as opposed to potential problems with the framework.

Appendix B. Proof of Proposition 2

Proof First for any model in $\mathbb{M}_{S=1}$ due to the ancestral closure and the fact $U \perp A$ we must have for all the potential outcomes $X(a)$ that is generates:

$$X(a) \perp A \ , \forall a \in A$$

Hence if we have for some a, $X^*(a) \not\perp A \mid S = 1$ then no model in $\mathbb{M}_{S=1}$ can generate these $X^*(a)$.

Now if we have $X^*(a) \perp A \mid S = 1 \ , \forall a \in A$ then we construct a causal model $\bar{\mathcal{M}} \in \mathbb{M}_{S=1}$ with the correct counterfactuals by taking our U to be $\{X^*(a), \forall a \in A\}$ and simply $X = \sum \mathbb{I}(A = a)X(a)$. This clearly generates our data for $S = 1$, we have $U \perp A$ as $X(a) \perp A \mid S = 1 \ , \forall a \in A$. Finally this causal model trivially has the same counterfactuals as $\mathcal{M}^*$ ■

Appendix C. Proof of Proposition 3

Proof First we note the independence $X^*(a) \perp A \mid S = 1$ is equivalent to saying for all a, a' with $P(S = 1 \mid A = a) > 0$ and $(S = 1 \mid A = a') > 0$:

$$P(X^*(a) = x \mid A = a, S = 1) = P(X^*(a) = x \mid A = a', S = 1).$$

Applying Bayes rule gives:

$$\frac{P(S = 1 \mid X^*(a) = x, A = a)P(X^*(a) = x \mid A = a)}{P(S = 1 \mid A = a)} = \frac{P(S = 1 \mid X^*(a) = x, A = a')P(X^*(a) = x \mid A = a')}{P(S = 1 \mid A = a')}.$$

Now we use the fact that $X^*(a) \perp A$ in the population, so we have $P(X^*(a) = x \mid A = a) = P(X^*(a) = x \mid A = a')$; hence since we have $P(X^*(a) = x) = P(X = x \mid A = a) > 0$ we can cancel these to give:

$$\frac{P(S = 1 \mid X^*(a) = x, A = a)}{P(S = 1 \mid A = a)} = \frac{P(S = 1 \mid X^*(a) = x, A = a')}{P(S = 1 \mid A = a')}.$$

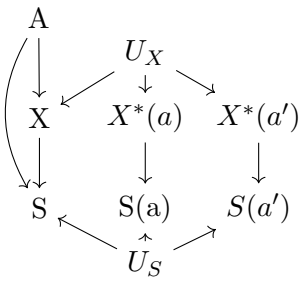
All that remains to show is that:

$$P(S = 1 \mid X^*(a) = x, A = a') = P(S(a') = 1 \mid X = x, A = a).$$

We have:

$$\begin{aligned} P(S = 1 \mid X^*(a) = x, A = a') &= P(S(a') = 1 \mid X^*(a) = x, A = a') && (1) \\ &= P(S(a') = 1 \mid X^*(a) = x, A = a) && (2) \\ &= P(S(a') = 1 \mid X = x, A = a), && (3) \end{aligned}$$

where (1) and (3) follow from the consistency property, and (2) follows from the fact that $S(a') \perp A \mid X^*(a)$. This can be read off the following 'triplet' network which is Markovian under our Assumption 1.



Hence under assumption one the required equation is equivalent to ignorability under selection. ■

Appendix D. Proof of Corollary 5

Proof By rearranging the requirement given in Proposition 3, we have that if $X^*(a) \perp A \mid S = 1$ then for all x with $P(X = x \mid A = a) > 0$:

$$\begin{aligned} P(S = 1 \mid X = x, A = a) &= P(S(a) = 1 \mid X = x, A = a)\frac{P(S = 1 \mid A = a)}{P(S = 1 \mid A = a')} \\ &= P(S(a) = 1 \mid X = x, A = a)\frac{P(A = a \mid S = 1)P(A = a')}{P(A = a' \mid S = 1)P(A = a)} \quad (4) \\ &\leq \frac{P(A = a \mid S = 1)P(A = a')}{P(A = a' \mid S = 1)P(A = a)} \quad (5) \end{aligned}$$

where (4) follows from applying Bayes' rule and (5) uses that a probability is bounded above by 1. Hence if there exists an x which violates this bound we must have $X^*(a) \not\perp A \mid S = 1$ and so by Lemma 2 no model in $\mathcal{M}_{S=1}$ captures the true counterfactuals. ■

Appendix E. Detailing the Calculations and Datasets

In this appendix we detail all the datasets and calculations from Table 1. All population data for this section is from the World Bank.

E.1. Adult Dataset

The Adult Dataset (Kohavi and Becker, 1994) contains data on 48 842 individuals taken from the 1994 US census database. The dataset contains 16 attributes for each individual with an aim to predict if an individuals income is greater than $50,000. The dataset was formed by taking all individuals in the US census database with these 16 attributes recorded and then removing based on certain attributes to get clean records. For example all individuals who were logged as not working any hours were removed.

In the this dataset the gender distribution is 67% male and 33% female. The World Bank estimates that in 1994, 49.1% of the US population were male and 50.9% were female.

Plugging this in gives:

$$\begin{aligned} P(S=1 \mid X=x, A=\text{Female}) &> \frac{P(A=\text{Female} \mid S=1)P(A=\text{Male})}{P(A=\text{Male} \mid S=1)P(A=\text{Female})} \\ &= \frac{0.33 \times 0.491}{0.67 \times 0.509} \\ &= 0.475. \end{aligned}$$

E.2. German Credit Dataset

The German Credit Dataset (Hofmann, 1994) describes has the financial details of 1000 bank customers applying for a loan. The task is to predict from a list of 20 covariates if someone is a good or bad credit risk. This dataset is from the year 1994.

In the German credit dataset the gender is 31% female and 69% male. The World Bank estimates that in 1994, 51.5% of the German population were female and 48.4% were male. This gives:

$$\begin{aligned} P(S=1 \mid X=x, A=\text{Female}) &> \frac{P(A=\text{Female} \mid S=1)P(A=\text{Male})}{P(A=\text{Male} \mid S=1)P(A=\text{Female})} \\ &= \frac{0.31 \times 0.484}{0.69 \times 0.516} \\ &= 0.421. \end{aligned}$$

E.3. Law School Dataset

The law Sshool dataset (Wightman, 1998) is as described in Section 2.2. The dataset was collected in 1998.

Again using World Bank estimates we have that in 1998 the US population is 49.7% female and 50.3% male. In the law school dataset the gender distribution is 43.8% female and 56.2% male. This gives:

$$\begin{aligned} P(S=1 \mid X=x, A=\text{Female}) &> \frac{P(A=\text{Female} \mid S=1)P(A=\text{Male})}{P(A=\text{Male} \mid S=1)P(A=\text{Female})} \\ &= \frac{0.438 \times 0.503}{0.562 \times 0.497} \\ &= 0.789. \end{aligned}$$

Appendix F. Proof of Lemma 6

Proof As noted previously we have $X(a) \perp A$ for the counterfactuals generated by causal models satisfying the assumptions on $\mathcal{M}$ in this Lemma. Hence as $\widehat{Y}$ is a function of X

(possibly also with some independent noise) we have $\widehat{Y}(a) \perp A$.

$$
\begin{aligned}
P(\widehat{Y} \mid A = a') &= P(\widehat{Y}(a') \mid A = a') && (6)\\
&= P(\widehat{Y}(a') \mid A = a) && (7)\\
&= \mathbb{E}_{P(X|A=a)}\Big(P(\widehat{Y}(a') \mid X, A = a)\Big) && (8)\\
&= \mathbb{E}_{P(X|A=a)}\Big(P(\widehat{Y}(a) \mid X, A = a)\Big) && (9)\\
&= P(\widehat{Y}(a) \mid A = a)\\
&= P(\widehat{Y} \mid A = a)
\end{aligned}
$$

Where (6) follows from consistency, (7) follows from $\widehat{Y}(a') \perp A$, (8) uses the law of total expectation and (9) uses the definition of counterfactual fairness. Hence $\widehat{Y} \perp A$. ■

Appendix G. Proof of Corollary 7

Proof We take predictors to mean any function, since counterfactual fairness places no restriction on what value the predictors take.

First if we satisfy the counterfactual outcome independence under selection then we have by Lemma 2 that we have a model in $\mathcal{M}_{S=1}$ with the correct counterfactuals and $\widehat{Y}$ is clearly counterfactually fair relative to this as it has the same counterfactuals as the true model. Therefore by Lemma 6 $\widehat{Y}$ is independent of A on the dataset, so when $S = 1$.

Now for the converse we show that functions f counterfactually fair according to $\mathcal{M}^*$ will not in general satisfy $f(X, A) \perp A \mid S = 1$ by finding a specific function which violates this.

First as $X^*(a) \not\perp A \mid S = 1$ for some a we have some coefficient X_1 such that $X_1^*(a) \not\perp A \mid S = 1$. Now let f_1 be a function such that for inputs $X = x$ and $A = a'$, $f_1(x, a')$ will be a random draw from the true posterior for $X^*(a)$ arising from $\mathcal{M}^*$, that is $P(X_1^*(a) \mid X = x, A = a')$.

Now, clearly this function will be counterfactually fair according to the true model. However we have $f_1(X, A) \not\perp A \mid S = 1$. This is because given $A = a'$ a random draw from $f_1(X, A) \mid \{S = 1\}$ will be a random draw from $X_1^*(a) \mid \{A = a', S = 1\}$. As we know $X_1^*(a) \not\perp A \mid S = 1$ we conclude $f_1(X, A) \not\perp A \mid S = 1$ and so we are done. ■

References

Professor Dr. Hans Hofmann. UCI machine learning repository, 1994. URL https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

Lily Hu and Issa Kohler-Hausmann. What's sex got to do with fair machine learning? *arXiv preprint arXiv:2006.01770*, 2020.

Atoosa Kasirzadeh and Andrew Smart. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 228–236, 2021.

Ronny Kohavi and Barry Becker. UCI machine learning repository, 1994. URL https://archive.ics.uci.edu/ml/datasets/adult.

Issa Kohler-Hausmann. Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*, 113:1163, 2018.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30, 2017.

Maya Sen and Omar Wasow. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19:499–522, 2016.

Linda F Wightman. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. 1998.

World Bank. World bank indicators. URL https://data.worldbank.org/indicator.

Edward N Zalta, Uri Nodelman, Colin Allen, and John Perry. *Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Center for the Study of Language and Information . . . , 1995.