# A Take on Obfuscation with Ethical Adversaries

**Pieter Delobelle**[*] PIETER.DELOBELLE@KULEUVEN.BE
**Paul Temple**[**] PAUL.TEMPLE@UNAMUR.BE
**Gilles Perrouin**[**] GILLES.PERROUIN@UNAMUR.BE
**Benoît Frénay**[**] BENOIT.FRENAY@UNAMUR.BE
**Patrick Heymans**[**] PATRICK.HEYMANS@UNAMUR.BE
**Bettina Berendt**[*&***] BETTINA.BERENDT@KULEUVEN.BE

[*] *Department of Computer Science, KU Leuven and Leuven.ai*

[**] *PReCISE, NaDi, Université de Namur*

[***] *TU Berlin; Weizenbaum Institute; KU Leuven*

In recent years, an increasing number of real-world observations and research findings have shown that machine learning (ML) can lead to *unfair* stereotyping and differences in outcomes between social groups. This is an issue for everything from small (Angwin and Larson 2016, Chouldechova 2017) to extremely large datasets (Blodgett et al. 2020).

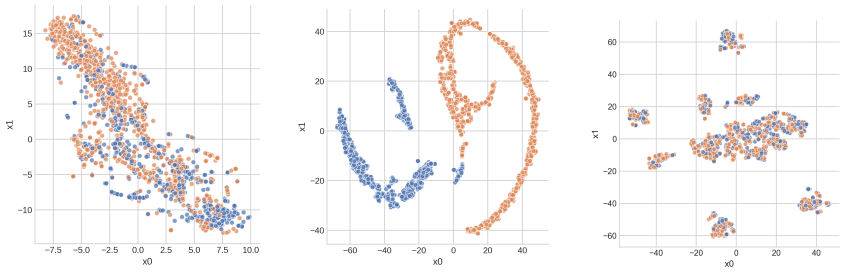
Although specific patterns of 'unfair' behaviours differ widely between fields (stereotypical embeddings in natural language processing, unequal performances in computer vision, etc.), a simple and general approach consists in ignoring protected attributes during training. However, this is not always straightforward and because of dependencies between features this approach provides few guarantees and may even lead to internal representations that could reconstruct protected attributes' value. This can significantly affect social groups (Angwin and Larson 2016, Corbett-Davies and Goel 2018, Pierson et al. 2018). In other cases, unfair behaviours stem from insufficient coverage of social groups in the training data, e.g. facial recognition algorithms with ethnicity-dependent performance.

To mitigate these systemic problems, we argue that *appropriate* obfuscation of protected attributes (such as gender or 'race') should be an integral part of machine learning. We demonstrate our point through our *Ethical Adversaries* framework (Delobelle et al. 2020), that was originally developed to incorporate such fairness constraints.

## 1. Ethical Adversaries for obfuscation

We previously proposed a method leveraging interactions between two adversaries (Delobelle et al. 2020): (i) an adversarial *feeder* that crafts adversarial examples and (ii) a *reader* that uses *Gradient Reversal Layer* (GRL) to both predict a protected attribute from input data and mitigate this ability in the model. Gradient reversal was introduced for domain adaptation (Ganin et al. 2016) and was used to create 'fair' machine learning models by viewing a protected attribute as a domain label. If the GRL of the model is unable to predict the protected attribute—or originally the domain label, then the predictions for the main model are considered fair (Raff and Sylvester 2018, Adel et al. 2019).

The feeder introduces an obfuscation strategy for the training inputs by generating counterfactual inputs that align all representations, irregardless whether an individual belongs to underrepresented groups or not. Indeed, the reader alone is insufficient, as internal representations would form clusters based on the protected attribute if used alone, see Figure 1 and Delobelle et al. (2020) for a more in-depth analysis. However, combining the reader with obfuscated training examples generated by the feeder mitigates this limitation: While Figure 1a shows a low-dimensionality representaiton of the data where example of both 'race' are mixed together, when applying GRL alone the groups become distinguishable (see Figure 1b), which can favor differ-



(a) Naive model (b) Model trained with a GRL ($\lambda = 50$) (c) Model trained with our framework

Figure 1: T-SNE dimensionality reduction of the activations in the last hidden layer on the held-out COMPAS test set. Distinct colors are used for the reported race of individuals in the dataset: either African-American ● or Caucasian ● . Originally published in Delobelle et al. (2020).

ent outcomes from the ML model. After applying Ethical adversaries (see Figure 1c), the visualisation shows that the groups are mixed together again and it is even harder to differentiate than with a naive model.

In the Ethical Adversaries framework, the obfuscation strategy targets the classification outputs of the reader. This strategy uses adversarial machine learning (Biggio and Roli 2018) to craft adversarial examples that reduce the model's ability to predict the protected attribute. Although the adversarial examples are used during training, they are generated by running an evasion attack (Biggio et al. 2013). Ultimately, the generation of adversarial examples along with a retraining of the classifier helps to hide protected attributes from the learning algorithm which corresponds to our obfuscation strategy.

The framework also allows practitioners to prioritise either the utility of the model or the reconstruction of the protected attribute. In the obfuscation literature, this is modelled as a trade-off between quality of service and privacy protection (Shokri et al. 2016)

## 2. Related work

There is a wide range of work on incorporating fairness constraints in learning algorithms (Calders and Žliobaitė 2013, Barocas et al. 2019) and on pre-processing training data. Specifically the work by Solans et al. (2020) is related to the feeder in our framework, where the authors use poisoning attacks to improve fairness.

When looking at the broader field, the work by Kulynych et al. (2020) on Protective Optimization Technologies (POTs) uses poisoning attacks as well to model user interactions as a defence strategy. The most relevant work can be found in the obfuscation literature, namely a framework by Romanelli et al. (2020). This framework uses a GAN-like (Goodfellow et al. 2014) model to remove sensitive attributes, like location data, and models this as a zero-sum game between a *leader* and *follower*. Although details in the implementation differ and the focus is on removing identifiable information, the parallels with our work do highlight a link between fairness and obfuscation.

## 3. Conclusion

In this work, we described Ethical Adversaries, which relies on obfuscation of protected attributes to reduce impact of ML decision on (protected) socials groups. Although our work was not developed within an obfuscation framework, the feeder can be modelled as such.

More interestingly are parallels that exist between fairness and obfuscation literature, not only with our work. We highlighted this with related work that yielded similar model designs, but such parallels also exist in problem formulation. It appears that "forgetting" or "hiding" certain information can be helpful in the quest for fairness (see also (Ruggieri 2014)), implying that obfuscation techniques are central tools for making machine learning fairer.

## Acknowledgements

Pieter Delobelle was supported by the Research Foundation - Flanders (FWO) under EOS No. 30992574 (VeriLearn). Pieter Delobelle also received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. Paul Temple is also supported by the EOS VeriLearn project (Fonds de la Recherche Scientifique, FNRS). Gilles Perrouin is an FNRS Research Associate. We also want to thank the secML developers from the PRALab (Pattern Recognition and Applications Laboratory, University of Cagliari, Sardegna, Italy) for having answered our numerous questions and helping us in using their newly developed library.

## References

Adel, Tameem, Isabel Valera, Zoubin Ghahramani, and Adrian Weller (2019), One-Network Adversarial Fairness, *AAAI Conference on Artificial Intelligence*.

Angwin, Julia and Jeff Larson (2016), Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks., *ProPublica*.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2019), *Fairness and Machine Learning*, fairmlbook.org.

Biggio, Battista and Fabio Roli (2018), Wild patterns: Ten years after the rise of adversarial machine learning, *Pattern Recognition Journal* **84**, pp. 317–331.

Biggio, Battista, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli (2013), Evasion attacks against machine learning at test time, *ECML/PKDD*, pp. 387–402.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach (2020), Language (Technology) is Power: A Critical Survey of "Bias" in NLP, *arXiv:2005.14050 [cs]*.

Calders, Toon and Indrė Žliobaitė (2013), Why unbiased computational processes can lead to discriminative decision procedures, *Discrimination and Privacy in the Information Society*, Springer, pp. 43–57.

Chouldechova, Alexandra (2017), Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, *arXiv:1703.00056*.

Corbett-Davies, Sam and Sharad Goel (2018), The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning, *arXiv:1808.00023*.

Delobelle, Pieter, Paul Temple, Gilles Perrouin, Benoît Frénay, Patrick Heymans, and Bettina Berendt (2020), Ethical adversaries: Towards mitigating unfairness with adversarial machine learning.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky (2016), Domain-adversarial training of neural networks, *The Journal of Machine Learning Research* **17** (1), pp. 2096–2030.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014), Generative Adversarial Nets, *NIPS*, Curran Associates, pp. 2672–2680.

Kulynych, Bogdan, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses (2020), POTs: Protective Optimization Technologies, *Proceedings of FAccT 2020* pp. 177–188.

Pierson, Emma, Sam Corbett-Davies, and Sharad Goel (2018), Fast threshold tests for detecting discrimination, *in* Storkey, Amos and Fernando Perez-Cruz, editors, *AISTAT 21*, Vol. 84 of *Proceedings of Machine Learning Research*, PMLR, Playa Blanca, Lanzarote, Canary Islands, pp. 96–105.

Raff, E. and J. Sylvester (2018), Gradient Reversal against Discrimination: A Fair Neural Network Learning Approach, *IEEE 5th International Conference on Data Science and Advanced Analytics*, pp. 189–198.

Romanelli, Marco, Kostas Chatzikokolakis, and Catuscia Palamidessi (2020), Optimal obfuscation mechanisms via machine learning, *2020 IEEE 33rd Computer Security Foundations Symposium (CSF)*, pp. 153–168.

Ruggieri, Salvatore (2014), Using t-closeness anonymity to control for non-discrimination, *Trans. Data Priv.* **7** (2), pp. 99–129. http://www.tdp.cat/issues11/abs.a196a14.php.

Shokri, Reza, George Theodorakopoulos, and Carmela Troncoso (2016), Privacy games along location traces: A game-theoretic framework for optimizing location privacy, *ACM Transactions on Privacy and Security (TOPS)* **19** (4), pp. 1–31, ACM New York, NY, USA.

Solans, David, Battista Biggio, and Carlos Castillo (2020), Poisoning attacks on algorithmic fairness, *arXiv preprint arXiv:2004.07401*.