A Variational Approach to Privacy and Fairness

Borja Rodríguez-Gálvez **Ragnar Thobaben** **Mikael Skoglund**
Division of Information Science and Engineering (ISE)
KTH Royal Institute of Technology
{borjarg,ragnart,skoglund}@kth.se

Abstract

In this article, we propose a new variational approach to learn private and/or fair representations. This approach is based on the Lagrangians of a new formulation of the privacy and fairness optimization problems that we propose. In this formulation, we aim at generating representations of the data that keep a prescribed level of the relevant information that is not shared by the private or sensitive data, while minimizing the remaining information they keep. The proposed approach (i) exhibits the similarities of the privacy and fairness problems, (ii) allows us to control the trade-off between utility and privacy or fairness through the Lagrange multiplier parameter, and (iii) can be comfortably incorporated to common representation learning algorithms such as the VAE, the β-VAE, the VIB, or the nonlinear IB.

1 Introduction

Currently, many systems rely on machine learning algorithms to make decisions and draw inferences. That is, they use previously existing data in order to shape some stage of their decision or inference mechanism. Usually, this data contains private or sensitive information, e.g., the identity of the person from which a datum was collected or their membership to a minority group. Therefore, an important problem occurs when the data used to train such algorithms leaks this information to the system, thus contributing to unfair decisions or to a privacy breach.

When the content of the private or sensitive information is arbitrary and the task of the system is not defined, the problem is reduced to learning *private representations* of the data; i.e., representations that are informative of the data (utility), but are not informative of the private or sensitive information. Then, these private representations can be employed by any system with a controlled leakage of private information. If the level of informativeness is measured with the mutual information, the problem of generating private representations is known as the *privacy funnel* (PF) [1, 2].

When the task of the system is known, then the aim is to design strategies so that the system performs such a task efficiently while obtaining or leaking little information about the sensitive data. The field of *algorithmic fairness* has extensively studied this problem, especially for classification tasks and categorical sensitive data; see, e.g., [3, 4, 5, 6, 7]. An interesting approach is that of learning *fair representations* [8, 9, 10], where similarly to their private counterparts, the representations are informative of the task output, but are not informative of the sensitive information.

There is a compromise between the information leakage and the utility when desigining private representations [2]. Similarly, in the field of algorithmic fairness, it has been shown empirically [11] and theoretically [12] that there is a trade-off between fairness and utility.

In this work, we investigate the trade-off between utility and privacy and between utility and fairness in terms of mutual information. More specifically, we aim at maintaining a certain level of the information about the data (for privacy) or the task output (for fairness) that is not shared by the senstive data, while minimizing all the other information. We name these two optimization problems the *conditional privacy funnel* (CPF) and the *conditional fairness bottleneck* (CFB) due to their

Preprint. Under review.

similarities with the privacy funnel [2], the information bottleneck [13], and the recent conditional entropy bottleneck [14].

We tackle both optimization problems with a variational approach based on their Lagrangian. For the privacy problem, we show that the minimization of the Lagrangians of the CPF and the PF is equivalent (see the supplementary material A.2), meaning that our variational approach attempts at solving the PF as well. Moreover, this approach improves over current variational approaches to the PF by respecting the problem's Markov chain in the encoder distribution.

Finally, the resulting approaches for privacy and fairness can be implemented with little modification to common algorithms for representation learning like the variational autoencoder (VAE) [15], the β-VAE [16], the variational information bottleneck (VIB) [17], or the nonlinear information bottleneck [18]. Therefore, it facilitates the incorporation of private and fair representations in current applications (see the supplementary material B for a guide on how to modify these algorithms).

We demonstrate our results both in the Adult dataset (available at [19]), which is commonly employed for benchmarking both tasks, and a high-dimensional toy dataset, based on the MNIST hand-written digits dataset [20]. Further experiments can be found in the supplementary material D.

2 Methods

In this section we give an overview of our approach. First, we introduce our proposed model for the privacy and fairness problems. Then, we present a suitable Lagrangian formulation. Finally, we describe a variational approach to solving both problems.

2.1 Problem formulation

2.1.1 Privacy formulation: the conditional privacy funnel (CPF)

Let us consider two random variables $X \in \mathcal{X}$ and $S \in \mathcal{S}$ such that $I(X;S) \geq 0$. The random variable X represents some data which is of interest to us, and the random variable S represents some private data. We wish to disclose the data of interest X; however, we do not want the receiver of this data to draw inferences about the private data S. For this reason, we encode the data of interest X into the representation $Y \in \mathcal{Y}$, forming the Markov chain $S \leftrightarrow X \rightarrow Y$.

This encoding, characterized by the conditional probability distribution $P_{Y|X}$, is designed so that the representation Y keeps a certain level r of the information about the data of interest X that is not shared by the private data S (i.e., the light gray area in Figure 1a), while minimizing the information it keeps about the private data S (i.e., the dark gray area in Figure 1a). That is,

$$\arg \inf_{P_{Y|X}} \{I(S;Y)\} \text{ s.t. } I(X;Y|S) \geq r. \tag{1}$$

The main difference with the privacy funnel formulation [2] is that, even though both formulations minimize the information the representation Y keeps about the private data S, in the PF the encoding is designed so that Y keeps a certain level r' of the information about X, regardless if this information is also shared by S. Hence, since $I(X;Y|S) \leq I(X;Y)$, the restrictions of the CPF on the representations are stronger, unless $X \perp S$, in which case $I(S;Y) = 0$ and they are equal. Nevertheless, the minimization of the Lagrangians of both problems is equivalent (see the supplementary material A.2).

2.1.2 Fairness formulation: the conditional fairness bottleneck (CFB)

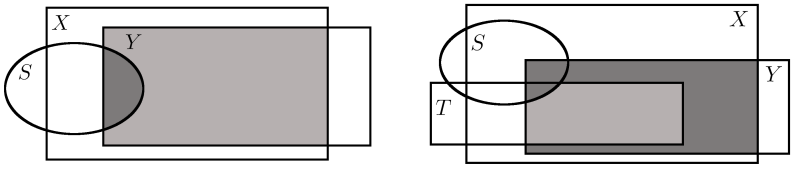
Let us consider three random variables $X \in \mathcal{X}, S \in \mathcal{S}$, and $T \in \mathcal{T}$ such that $I(X;S) \geq 0$ and $I(S;T) \geq 0$. The random variable X represents some data we want to use to draw inferences about the task T. However, we do not want our inferences to be influenced by the sensitive data S. For this reason, we first encode the data X into a representation $Y \in \mathcal{Y}$, which is then used to draw inferences about T. Therefore, the Markov chains $S \leftrightarrow X \rightarrow Y$ and $T \leftrightarrow X \rightarrow Y$ hold.

This encoding, characterized by the conditional probability distribution $P_{Y|X}$, is designed so that the representation Y keeps a certain level r of the information about the task output T that is not shared by the private data S (i.e., the light gray area in Figure 1b), while minimizing both the information it keeps about the private data S and the information about X that is not shared with the task output T

(i.e., the dark gray area in Figure 1b). That is,

$$\arg \inf_{P_{Y|X}} \{I(S;Y) + I(X;Y|S,T)\} \text{ s.t. } I(T;Y|S) \geq r. \tag{2}$$

This formulation differs from other approaches to fairness mainly in two points: (i) Similarly to the information bottleneck [13], the CFB does not only minimize the information the representations Y keep about the sensitive data S, but also minimizes the information about X that is not relevant to draw inferences about T. That is, the CFB seeks representations that are both *fair* and *relevant*, thus avoiding the risk of keeping *nuisances* [21] and harming their generalization capability. (ii) Similarly to the conditional entropy bottleneck [14], the CFB aims to produce representations Y that keep a certain level r of the information about the task T that is not shared by S. This differs from formulations that aim at producing representations that maintain a certain level r' of the information about T, regardless if it is also shared by the sensitive data S.



(a) Conditional privacy funnel (b) Conditional information bottleneck

Figure 1: Information Diagrams [22] of (a) the conditional privacy funnel and (b) the conditional information bottleneck. In light gray, the relevant information we would like Y to keep. In dark gray, the useless information we would like Y to discard.

2.2 The Lagrangians of the problems

A common approach to solving optimization problems such as the CPF or the CFB is to minimize the *Lagrangian* of the problem. The Lagrangian is a proxy of the trade-off between the function to optimize and the constraints on the optimization search space [23, Chapter 5]. Particularly, the Lagrangians of the CPF and the CFB are, respectively,

$$\mathcal{L}_{\text{CPF}}(P_{Y|X}, \lambda) = I(S;Y) - \lambda I(X;Y|S) \text{ and} \tag{3}$$

$$\mathcal{L}_{\text{CFB}}(P_{Y|X}, \lambda) = I(S;Y) + I(X;Y|S,T) - \lambda I(T;Y|S), \tag{4}$$

where $\lambda > 0 \in \mathbb{R}$ is the *Lagrange multiplier* of the Lagrangian.[1] This multiplier controls the trade-off between the information the representations Y discard and the information they keep.

If we look at the information diagrams [22] of the CPF and the CFB from Figure 1, we observe how $\mathcal{L}_{\text{CPF}}(P_{Y|X}, \lambda)$ and $\mathcal{L}_{\text{CFB}}(P_{Y|X}, \lambda)$, by means of the multiplier λ, exactly control the trade-off between the information we want the representations to keep (i.e., the light gray area) and all the other information (i.e., the dark gray area).

In the following propositions, proved in the supplementary material A.1, we present two alternative Lagrangians that can be minimized instead of the original problem Lagrangians in order to obtain the same result. These Lagrangians are more tractable and exhibit similar properties and structure in the privacy and fairness problems.

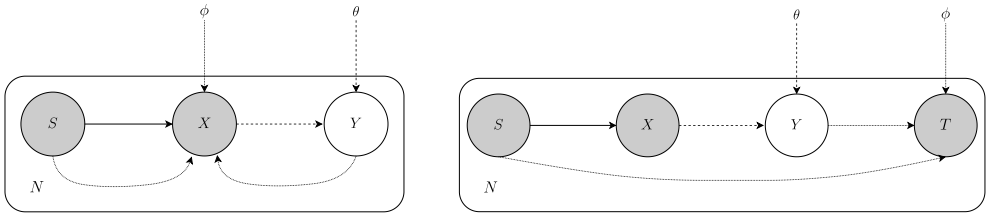
Proposition 1. *Minimizing $\mathcal{L}_{\text{CPF}}(P_{Y|X}, \lambda)$ is equivalent to minimizing $\mathcal{J}_{\text{CPF}}(P_{Y|X}, \gamma)$, where $\gamma = \lambda + 1$ and*

$$\mathcal{J}_{\text{CPF}}(P_{Y|X}, \gamma) = I(X;Y) - \gamma I(X;Y|S). \tag{5}$$

Proposition 2. *Minimizing $\mathcal{L}_{\text{CFB}}(P_{Y|X}, \lambda)$ is equivalent to minimizing $\mathcal{J}_{\text{CFB}}(P_{Y|X}, \beta)$, where $\beta = \lambda + 1$ and*

$$\mathcal{J}_{\text{CFB}}(P_{Y|X}, \beta) = I(X;Y) - \beta I(T;Y|S). \tag{6}$$

[1] Note that if $\lambda \leq 0$ the optimization only seeks for maximally compressed representations Y. Hence, trivial encoding distributions like a degenerate distribution $P_{Y|X}$ with density $p_{Y|X} = \delta(Y)$ are minimizers of the Lagrangian.



(a) Variational conditional privacy funnel (b) Variational conditional fairness bottleneck

Figure 2: Graphical models of (a) the variational conditional privacy funnel and (b) the variational conditional fairness bottleneck. The solid line represents the data density $p_{(S,X)}$. The dashed lines represent the encoding density $p_{Y|(X,\theta)}$ and the variational marginal density of the representations $q_{Y|\theta}$. The dotted lines represent (a) the generative variational density $q_{X|(S,Y,\phi)}$ or (b) the inference variational density $q_{T|(S,Y,\phi)}$. The encoding and the (a) generative or (b) inference parameters (i.e., θ and ϕ, respectively) are jointly learned.

We note that the minimization of $\mathcal{J}_{\text{CPF}}(P_{Y|X}, \gamma)$ and $\mathcal{J}_{\text{CFB}}(P_{Y|X}, \beta)$, by means of γ and β, trades off the level of compression of the representations with the information they keep, respectively, about X and T that is not shared by S.

2.3 The variational approach

We consider the minimization of $\mathcal{J}_{\text{CPF}}(P_{Y|X}, \gamma)$ and $\mathcal{J}_{\text{CFB}}(P_{Y|X}, \beta)$ to solve the CPF and the CFB problems. Furthermore, we assume that the probability density function[2] $p_{Y|X}$ that describes the conditional probability distribution $P_{Y|X}$ exists and it is parameterized by θ, i.e., $p_{Y|X} = p_{Y|(X,\theta)}$.

The Markov chains of the CPF and the CFB establish that the variables' joint densities are factored as $p_{(S,X,Y)|\theta} = p_{(S,X)}p_{Y|(X,\theta)}$ and $p_{(S,T,X,Y)|\theta} = p_{(S,T,X)}p_{Y|(X,\theta)}$, respectively. The densities $p_{(S,X)}$ and $p_{(S,T,X)}$ can be inferred from the data and the density $p_{Y|(X,\theta)}$ is to be designed.

The term $I(X;Y)$ depends on the density $p_{Y|\theta}$, which is usually intractable. Similarly, the terms $I(X;Y|S)$ and $I(T;Y|S)$ depend on the densities $p_{X|(S,Y,\theta)}$ and $p_{T|(S,Y,\theta)}$, respectively, which are also usually intractable. Therefore, an exact optimization of θ would be prohibitively computationally expensive. For this reason, we introduce the variational density approximations $q_{Y|\theta}$, $q_{X|(S,Y,\phi)}$, and $q_{T|(S,Y,\phi)}$, where the generative and inference densities are parametrized by ϕ. This variational approximation leads to the graphical models displayed in Figure 2. Then, as previously done in, e.g., [15, 24, 18], we leverage the non-negativity of the Kullback-Leibler divergence [25, Theorems 2.6.3 and 8.6.1] to bound $\mathcal{J}_{\text{CPF}}(P_{Y|X}, \gamma)$ and $\mathcal{J}_{\text{CFB}}(P_{Y|X}, \beta)$ from above. More precisely, we find an upper bound on $I(X;Y)$ and a lower bound on both $I(X;Y|S)$ and $I(T;Y|S)$, i.e.,

$$I(X;Y) = \mathbb{E}_{p_{(X,Y)|\theta}}\left[\log\left(\frac{p_{Y|(X,\theta)}}{p_{Y|\theta}}\right)\right] \leq \mathbb{E}_{p_X}\left[D_{\text{KL}}\left(p_{Y|(X,\theta)}||q_{Y|\theta}\right)\right], \tag{7}$$

$$I(X;Y|S) = \mathbb{E}_{p_{(S,X,Y)|\theta}}\left[\log\left(\frac{p_{X|(S,Y,\theta)}}{p_{X|S}}\right)\right] \geq \mathbb{E}_{p_{(S,X,Y)|\theta}}\left[\log\left(\frac{q_{X|(S,Y,\phi)}}{p_{X|S}}\right)\right], \text{ and} \tag{8}$$

$$I(T;Y|S) = \mathbb{E}_{p_{(S,T,Y)|\theta}}\left[\log\left(\frac{p_{T|(S,Y,\theta)}}{p_{T|S}}\right)\right] \geq \mathbb{E}_{p_{(S,T,Y)|\theta}}\left[\log\left(\frac{q_{T|(S,Y,\phi)}}{p_{T|S}}\right)\right]. \tag{9}$$

Finally, we can jointly learn θ and ϕ through gradient descent. First, we note that the terms $\mathbb{E}_{p_{(S,X)}}[\log p_{X|S}]$ and $\mathbb{E}_{p_{(S,T)}}[\log p_{T|S}]$ do not depend on the parametrization. Second, we leverage the *reparametrization trick* [15], which allow us to compute an unbiased estimate of the gradients. That is, we let $p_{Y|X}dY = p_E dE$, where E is a random variable and $Y = f(X, E; \theta)$ is a deterministic function. Lastly, we estimate $p_{(S,X)}$ and $p_{(S,T,X)}$ as the data's empirical densities.

[2]Note that if $|\mathcal{Y}|$ is countable the probabiltiy density function is the probability mass function.

Therefore, in practice, if we have a dataset $D = \left\{\left(x^{(i)}, s^{(i)}\right)\right\}_{i=1}^{N}$ for the CPF or a dataset $D = \left\{\left(x^{(i)}, s^{(i)}, t^{(i)}\right)\right\}_{i=1}^{N}$ for the CFB, we minimize, respectively, the following cost functions:[3]

$$\tilde{\mathcal{J}}_{\text{CPF}}(\theta, \phi, \gamma) = \frac{1}{N}\sum_{i=1}^{N} D_{\text{KL}}\left(p_{Y|(X=x^{(i)},\theta)}||q_{Y|\theta}\right) - \gamma\mathbb{E}_{p_E}\left[\log\left(q_{X|(S=s^{(i)},Y=f(x^{(i)},E),\phi)}(x^{(i)})\right)\right] \tag{10}$$

$$\tilde{\mathcal{J}}_{\text{CFB}}(\theta, \phi, \beta) = \frac{1}{N}\sum_{i=1}^{N} D_{\text{KL}}\left(p_{Y|(X=x^{(i)},\theta)}||q_{Y|\theta}\right) - \beta\mathbb{E}_{p_E}\left[\log\left(q_{T|(S=s^{(i)},Y=f(x^{(i)},E),\phi)}(t^{(i)})\right)\right]. \tag{11}$$

An *a posteriori* interpretation of this approach is that if the encoder compresses the representations Y assuming that the decoder will use both Y and the private or sensitive data S, then the encoder will discard the information about S contained in the original data X in order to generate Y.

Remark 1. *Note that the resulting cost functions for the CPF and the CFB ressemble those of the VAE [15], the β-VAE [16], the VIB [24], or the nonlinear IB [18]. Let us consider the (common) case that the decoder density is estimated with a neural network. If such a network is modified so that it receives as input both the representations and the private or sensitive data instead of only the representations, then the optimization of these algorithms results in private and/or fair representations (see Appendix B for the details).*

3 Related work

3.1 Privacy

If the secret information S is the identity of the samples or their membership to a certain group, the field of *differential privacy* provides a theoretical framework for defining privacy and several mechanisms able to generate privacy-preserving queries about the data X and explore such data, see, e.g., [26]. If, on the other hand, the secret information is arbitrary, the theoretical framework introduced in [1] is commonly adopted, with the special case of the privacy funnel [2] when the utility and the privacy are measured with the mutual information.

The original greedy algorithm to compute the PF [2] assumes the data is discrete or categorical and does not scale. For this reason, approaches that take advantage of the scalability of deep learning have emerged. For instance, in [9] they learn the representations through adversarial learning, while in the the privacy preserving variational autoencoder (PPVAE) [17] and the unsupervised version of the variational fair autoencoder (VFAE) [27] they learn such representations with variational inference.

At their core, the PPVAE and the unsupervised VFAE end up minimizing the cost functions

$$\mathcal{J}_{\text{PPVAE}}(\theta, \phi, \eta) = \mathbb{E}_{p_{(S,X)|\theta}}[D_{\text{KL}}(p_{Y|(S,X,\theta)}||q_{Y|\theta}] - \eta^{-1}\mathbb{E}_{p_{(S,X,Y)|\theta}}[\log(q_{X|(S,Y,\phi)}] \tag{12}$$

and $\mathcal{J}_{\text{VFAE}}(\theta, \phi, \delta) = \mathcal{J}_{\text{PPVAE}}(\theta, \phi, 1) + \delta\mathcal{J}_{\text{MMD}}(\theta, \phi)$, where $\mathcal{J}_{\text{MMD}}(\theta, \phi)$ is a maximum-mean discrepancy term. Even though the resulting function to optimize is similar to ours, it is important to note the encoding density in these works is $p_{Y|(S,X,\theta)}$, which does not respect the problem's Markov chain $S \leftrightarrow X \rightarrow Y$. Therefore, the optimization search space includes representations Y that contain information about the private data S that is not even contained in the original data X. Moreover, the private data S might not be available during inference.

3.2 Fairness

The field of algorithmic fairness is mainly dominated by the notions of *individual fairness*, where the sensitive data S is the identity of the data samples, and *group fairness*, where S is a binary variable that represents the membership of the data samples to a certain group. There are several approaches that aim at producing classifiers that ensure either of these notions of fairness; e.g., discrimination-free naive Bayes [11], constrained logistic regression, hinge loss and support vector machines [5], or regularized logistic regression through the Wasserstein distance [28].

[3] In fact, the expectation over $E \in \mathcal{E}$ is estimated with a naive Monte Carlo estimation of a single sample; i.e., the expectation of $g(E) : \mathcal{E} \mapsto \mathbb{R}$ is estimated as $\mathbb{E}_{p_E}[g(E)] \approx \frac{1}{L}\sum_{l=1}^{L} g(\epsilon^{(l)})$, where $\epsilon^{(l)} \sim p_E$ and $L = 1$.

Other lines of work on algorithmic fairness are based on causal inference [29, 30, 31, 32, 33, 34] and data massaging [35], where the values of the labels of the training data are changed so that the training data is fair.

The notion of fair representations, introduced by Zemel et al. [8], boosted the advances on algorithmic fairness due to the expressiveness of deep learning. These advances are mainly dominated by adversarial learning [8, 9, 10], even though there are recent variational approaches, too [27, 36].

The main difference with the variational approach from Creager et al. [36] is our simple cost function. They generate two representations, Y_{sens} and $Y_{\text{non-sens}}$, that contain the information about the sensitive data and the original data, respectively, necessary to draw inferences about the task T. At inference time, the sensitive representations Y_{sens} are corrupted with noise or discarded, and thus the non-sensitive representations $Y_{\text{non-sens}}$ from [36] serve a similar purpose to the representations Y obtained with our approach. Compared to the *variational fair autoencoder* [27], our encoding density does not require the sensitive information S, which might not be available during inference, thus not breaking the Markov chain $S \leftrightarrow X \rightarrow Y$.

4 Results

In this section, we present experiments on two datasets to showcase the performance of the presented variational approach to the privacy and fairness problems. First, we show the performance of the proposed method in a dataset commonly used for benchmarking both tasks. Second, we show the performance on high-dimensional data on a toy dataset especially designed for this purpose. The encoder density is modeled with an isotropic Gaussian distribution, i.e., $p_{Y|(X,\theta)} = \mathcal{N}(Y; \mu_{\text{enc}}(X;\theta), \sigma_\theta^2 I_d)$, so that $Y = \mu_{\text{enc}}(X;\theta) + \sigma_\theta \mathcal{N}(0, I_d)$, where μ_{enc} is a neural network and d is the dimensionality of the representations. The marginal density of the representations is also modeled as an isotropic Gaussian $q_{Y|\theta} = \mathcal{N}(Y; 0, I_d)$. Finally, the decoder density, $q_{X|(S,Y,\phi)}$ or $q_{T|(S,Y,\phi)}$, is modeled with a product of categorical (for discrete data) and/or isotropic Gaussians (for continuous data), e.g., $q_{X|(S,Y,\phi)} = \text{Cat}(X_1; \rho_{\text{dec}}(Y, S; \phi))\mathcal{N}(X_2; \mu_{\text{dec}}(S, Y; \phi), \sigma_\phi^2)$ if X consists of a discrete variable X_1 and a continuous variable X_2. The experiments are detailed in the supplementary material D.

Adult dataset The Adult dataset[4] contains $45,222$ samples from the 1994 U.S. Census. Each sample comprises 15 features such as, e.g., *gender*, *age*, *income level* (binary variable stating if the income level is higher or lower than $\$50,000$), or *education level*. For both tasks, we followed the experimental set-up from Zemel et al. [8] and considered S to be the gender and X to be the rest of the features. For the fairness task, we considered T to be the income level.

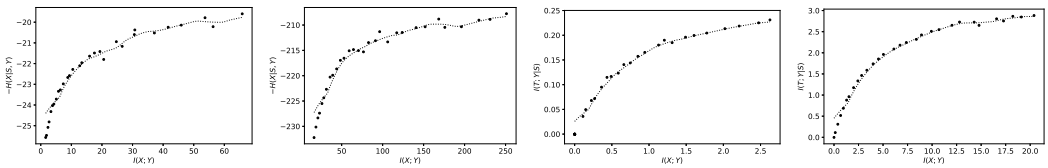
Toy dataset: Colored MNIST The MNIST dataset [20] is a collection of $70,000$ grayscale 28×28 images of hand-written digits from 0 to 9. The colored MNIST is a modification of the former dataset where each digit is randomly colored in either red, green, or blue. In both tasks we considered X to be the $3 \times 28 \times 28$ digit images, and S to be the color of the digit. Then, for the fairness task we considered T to be the digit number.

4.1 Privacy

In the privacy task, our proposed variational approach is able to control the trade-off between private and informative representations for both the Adult and the Colored MNIST datasets. We minimized (10) for different values of $\gamma \in [1, 50]$, thus controlling the trade-off between the compression level $I(X;Y)$ and the informativeness of the representations independent of the private data $I(X;Y|S)$ (or equivalently $-H(X|S,Y)$), as shown in Figures 3a and 3b. Therefore, as suggested by Proposition 1 and shown in Figures 4a and 4b, we also control the amount of information the representations keep about the private data, $I(S;Y)$, which was estimated with MINE [37].

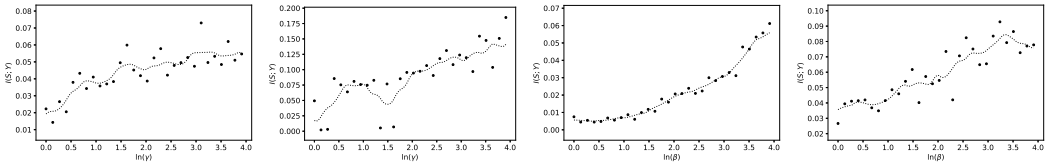
As an illustrative example, we constructed a representation of the same dimensionality, i.e., $3 \times 28 \times 28$, of the hand-written digits by minimizing (10) and setting $\gamma = 1$. This representation is both informative and private, as shown in Figures 5b and 5d. In Figure 5d, the 2-dimensional UMAP [38] vectors of the representations are mingled with respect to their color, as opposed to the UMAP vectors of the original images, where the vectors are clustered by the color of their images (see Figure 5c).

[4] Available at the UCI machine learning repository [19].



(a) (b) (c) (d)

Figure 3: Trade-off between the representations compression $I(X;Y)$ and the non-private information retained $I(X;Y|S)$ for the (a) Adult and the (b) colored MNIST datasets with $\gamma \in [1, 50]$. Since $I(X;Y|S) = H(X|S) - H(X|S,Y)$ and $H(X|S)$ does not depend on Y, the reported quantity is $-H(X|S,Y)$. Moreover, trade-off between the compression of the representations $I(X;Y)$ and the non-sensitive information retained about the task $I(T;Y|S)$ for the (c) Adult and the (d) colored MNIST datasets with $\beta \in [1, 50]$. The dots are the computed empirical values and the lines are the moving average of the 1D linear interpolations of such points.



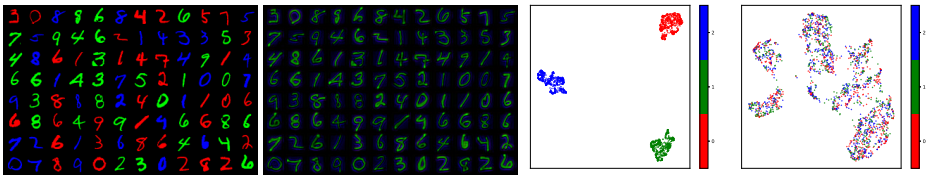
(a) (b) (c) (d)

Figure 4: Behavior of the private information $I(S;Y)$ kept by the private representations in the (a) Adult and (b) the colored MNIST datasets with $\gamma \in [1, 50]$; and by the fair representations in the (c) Adult and (d) the colored MNIST datasets with $\beta \in [1, 50]$. The dots are the computed empirical values and the lines are the moving average of the 1D linear interpolations of such points.

4.2 Fairness

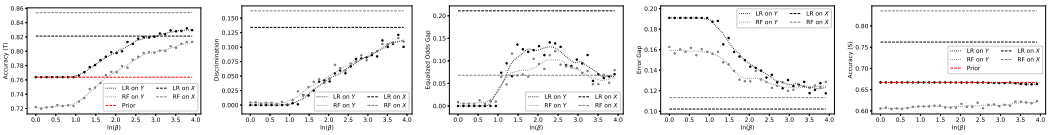
In the fairness task, our proposed variational approach is also able to control the trade-off between fair and accurate representations. We minimized (11) for different values of $\beta \in [1, 50]$, thus controlling the trade-off between the compression level $I(X;Y)$ and the predictability without the sensitive variable $I(T;Y|S)$, as shown in Figures 3c and 3d. Moreover, as suggested by Proposition 2 and shown in Figures 4c and 4d, we also control the amount of information that the representations retain about the sensitive data $I(S;Y)$, which was estimated with MINE [37].

Furthermore, in the Adult dataset, the Lagrange multiplier β allows us to control the behavior of different utility and group fairness indicators (defined in the supplementary material E), namely the accuracy, the error gap, and the discrimination. That is, the higher the value of β, the higher the accuracy and the discrimination, and the lower the error gap (Figures 6a, 6b, and 6d). The behavior



(a) (b) (c) (d)

Figure 5: Colored MNIST (a) original data, (b) private representations, (c) original data UMAP dimensionality reduction, and (d) private representations UMAP dimensionality reduction. In the UMAP dimensionality reduction, each vector point is colored with the same color the digit was. Results obtained for $\gamma = 1$.



(a) (b) (c) (d) (e)

Figure 6: Behavior of (a) the accuracy on T, (b) the discrimination, (c) the equalized odds gap, (d) the error gap, and (e) the accuracy on S of a logistic regression (LR, in black) and a random forest (RF, in gray) of the fair representations (dots and dotted line) and the original data (dashed line) learned with $\beta \in [1, 50]$ on the Adult dataset. The dashed line in red is the accuracy of a prior-based classifier.

of the discrimination is enforced by the minimization of $I(S;Y)$, as discussed in the supplementary material E. However, there is no clear indication of the effect of β on the accuracy of adversarial predictors on the sensitive data (which is still below the prior probability of the biased training dataset) and on the equalized odds (Figures 6e and 6c).

5 Discussion

In this article, we studied the problem of mitigating the leakage of private or sensitive information S into data-driven systems through the training data X. We formalized the trade-off between the relevant information for the system that is not shared by the private or sensitive data S and the remaining information as a constrained optimization problem. When the task of the system is known, the problem is referred to as learning fair representations and the formalization is the conditional fairness bottleneck (CFB); and when the task is unknown, the problem is referred to as learning private representations and the formalization is the conditional privacy funnel (CPF).

We proposed a variational approach, based on the Lagrangians of the CPF and the CFB, to solve the problem. This approach leads to a simple structure where the tasks of learning private and/or fair representations can be easily identified. Moreover, in practice, private and fair representations can be learned with little modification to the implementation of common algorithms such as the VAE [15], the β-VAE [16], the VIB [24], or the nonlinear IB [18]. Namely, modifying the decoder neural network so it receives both the representation Y and the private or sensitive data S as an input. Then, the learned representations can be fed to any algorithm of choice. For this reason, the efforts for reducing unfair decisions and privacy breaches will be small for many practitioners.

5.1 Limitations and future directions

Problem formulation The CPF and the CFB as defined in (1) and (2) are non-convex optimization problems with respect to $P_{Y|X}$. More specifically, they are a minimization of a convex function with non-convex constraints (see Appendix C). Therefore, (i) the optimal conditional distribution $P^{\star}_{Y|X}$ that minimizes the Lagrangian might not be achieved through gradient descent, and (ii) even if $P^{\star}_{Y|X}$ is achieved, it could be a sub-optimal value for (1) or (2), since the problems are not *strongly dual* [23, Section 5.2.3]. A possible solution could be the application of a monotonically increasing concave function u to $I(X;Y|S)$ or $I(T;Y|S)$ in the CPF or CFB Lagrangians, respectively, so that $u(I(X;Y|S))$ or $u(I(T;Y|S))$ is concave (and hence the Lagrangian is convex) in the domain of interest. For some u, this approach might allow to attain the desired r in (1) or (2) with a specific value of the Lagrange multiplier; see [39] for an example of this approach for the information bottleneck.

Proposed approach The proposed approach entails two limitations that are common in variational attempts at solving an optimization problem. Namely: (i) it approximates the decoding and the marginal distributions and (ii) it considers parametrized densities. The first issue restricts the search space of the possible encoding distributions $P_{Y|X}$ to those distributions with a decoding and marginal distributions that follow the restrictions of the variational approximation. The second issue further limits the search space to the obtainable encoding distributions with densities $p_{Y|(X,\theta)}$ with a parametrization θ. For this reason, richer encoding distributions and marginals, e.g., by means of normalizing flows [40, 41], are a possible direction to mitigate these issues.

Broader impact

Privacy breaches and unfairness are concerning problems that often arise in (learning) algorithms [42, 43]. Moreover, aside from the realm of algorithms, these are problems that humans, as a society, aim to mitigate in our objective to reach sustainability [44, Goal 10]. With the present paper, we contribute towards the development of privacy-preserving and fair systems, thus contributing to a sustainable development. For example, fair decision-making algorithms could directly contribute to the UN's targets 10.2 and 10.3 [44], namely, "empower and promote the social, economic, and political inclusion of all, irrespective of age, sex, disability, race, ethnicity, origin, religion, or economic or other status" and " ensure equal opportunity and reduce inequalities of outcome, including by eliminating discriminatory laws, policies, and practices and promoting appropriate legislation, policies, and action in this regard", respectively.

Even though our contribution will not help to directly level out social, political, and economic inequalities, the results and algorithms provided in this paper will help to avoid that inequalities will be amplified and prolonged through data-driven services and decision mechanisms (e.g., for insurances, administration, banking, or loans). By treating the data entered into such services and systems fairly and confidentially, as enforced by the proposed approach of this paper, our contribution has the potential to empower and promote social and economical inclusion [44, Target 10.2] and ensure equal opportunity [44, Target 10.3] in this specific domain of people's everyday life. Furthermore, we believe that our results are likely to be adopted by algorithm designers and practitioners, as our solutions can easily be integrated into existing standard representation learning algorithms as noted in Remark 1 and detailed in the supplementary material B.

Disclosure of funding

This work was supported in part by the Swedish Research Council.

References

[1] Flávio du Pin Calmon and Nadia Fawaz. Privacy against statistical inference. In *Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1401–1408. IEEE, 2012.

[2] Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Médard. From the information bottleneck to the privacy funnel. In *IEEE Information Theory Workshop (ITW)*, pages 501–505. IEEE, 2014.

[3] Toon Calders and Sicco Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[4] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science conference*, pages 214–226, 2012.

[5] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez Gomez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[6] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[7] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[8] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning (ICML)*, pages 325–333, 2013.

[9] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *International Conference on Learning Representations (ICLR)*, 2016.

[10] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. Conditional learning of fair representations. *International Conference on Learning Representations (ICLR)*, 2020.

[11] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

[12] Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 15649–15659, 2019.

[13] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[14] Ian Fischer. The conditional entropy bottleneck. *arXiv preprint arXiv:2002.05379*, 2020.

[15] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[16] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations (ICLR)*, 2017.

[17] Lihao Nan and Dacheng Tao. Variational approach for privacy funnel optimization on continuous data. *Journal of Parallel and Distributed Computing*, 137:17–25, 2020.

[18] Artemy Kolchinsky, Brendan D Tracey, and David H Wolpert. Nonlinear information bottleneck. *Entropy*, 21(12):1181, 2019.

[19] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[21] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

[22] Raymond W Yeung. A new outlook on Shannon's information measures. *IEEE Transactions on Information Theory*, 37(3):466–474, 1991.

[23] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[24] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *International Conference on Learning Representations (ICLR)*, 2016.

[25] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, second edition edition, 2006.

[26] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[27] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *International Conference on Learning Representations (ICLR)*, 2016.

[28] Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

[29] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 656–666, 2017.

[30] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4066–4076, 2017.

[31] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[32] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[33] Yixin Wang, Dhanya Sridhar, and David M Blei. Equal opportunity and affirmative action via counterfactual predictions. *arXiv preprint arXiv:1905.10870*, 2019.

[34] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 349–358, 2019.

[35] Faisal Kamiran and Toon Calders. Classification with no discrimination by preferential sampling. In *Proceedings of the 19th Machine Learning Conference of Belgium and The Netherlands*, pages 1–6. Citeseer, 2010.

[36] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning (ICML)*, pages 1436–1445, 2019.

[37] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning (ICML)*, pages 531–540, 2018.

[38] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29), 2018.

[39] Borja Rodríguez Gálvez, Ragnar Thobaben, and Mikael Skoglund. The convex information bottleneck Lagrangian. *Entropy*, 22(1):98, 2020.

[40] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

[41] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4743–4751, 2016.

[42] Songül Tolan. Fair and unbiased algorithmic decision making: Current state and future challenges. Technical report, European Commission, Joint Research Centre (JRC), 2018.

[43] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[44] United Nations (UN). Transforming our world: The 2030 agenda for sustainable development. 2016.

[45] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[46] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, volume 33, pages 5885–5892, 2019.

[47] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning (ICML)*, pages 2649–2658, 2018.

[48] Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2610–2620, 2018.

[49] William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 2016.

[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[51] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3315–3323, 2016.

Supplementary material

A Equivalences of the Lagrangians

In this section of the supplementary material, we show how minimizing the Lagrangians of the CPF and the CFB problems is equivalent to minimizing other Lagrangians. First, in A.1 we show it is equivalent to minimizing the Lagrangians that are used in the variational approach we propose in this paper. Then, in A.2 we show that minimizing the Lagrangian of the CPF is equivalent to minimizing the Lagrangian of the PF, meaning that the conditional probability distributions $P_{Y|X}$ obtained using the Lagrangian of the CPF would have been obtained through the Lagrangian of the PF, too.

A.1 Equivalence of the Lagrangians used for the minimization

Proposition 1. *Minimizing $\mathcal{L}_{\text{CPF}}(P_{Y|X}, \lambda)$ is equivalent to minimizing $\mathcal{J}_{\text{CPF}}(P_{Y|X}, \gamma)$, where $\gamma = \lambda + 1$ and*

$$\mathcal{J}_{\text{CPF}}(P_{Y|X}, \gamma) = I(X;Y) - \gamma I(X;Y|S). \tag{5}$$

Proof. If we manipulate the expression of the CPF Lagrangian we can see how minimizing $\mathcal{L}_{\text{CPF}}(P_{Y|X}, \lambda)$ is equivalent to minimizing $\mathcal{J}_{\text{CPF}}(P_{Y|X}, \gamma)$, where $\gamma = \lambda + 1$. More specifically,

$$\begin{aligned}
\arg\inf_{P_{Y|X}\in\mathcal{P}} \left\{\mathcal{L}_{\text{CPF}}(P_{Y|X}, \lambda)\right\} &= \arg\inf_{P_{Y|X}\in\mathcal{P}} \left\{I(S;Y) - \lambda I(X;Y|S)\right\} && (13)\\
&= \arg\inf_{P_{Y|X}\in\mathcal{P}} \left\{I(X;Y) - I(X;Y|S) - \lambda I(X;Y|S)\right\} && (14)\\
&= \arg\inf_{P_{Y|X}\in\mathcal{P}} \left\{I(X;Y) - (\lambda+1) I(X;Y|S)\right\} && (15)\\
&= \arg\inf_{P_{Y|X}\in\mathcal{P}} \left\{\mathcal{J}_{\text{CPF}}(P_{Y|X}, \lambda+1)\right\}, && (16)
\end{aligned}$$

where $\mathcal{P}$ is the set of probability distributions over $\mathcal{Y}$ such that if $P_{Y|X} \in \mathcal{P}$ for all $X \in \mathcal{X}$, then the Markov chain $S \leftrightarrow X \rightarrow Y$ holds. □

Proposition 2. *Minimizing $\mathcal{L}_{\text{CFB}}(P_{Y|X}, \lambda)$ is equivalent to minimizing $\mathcal{J}_{\text{CFB}}(P_{Y|X}, \beta)$, where $\beta = \lambda + 1$ and*

$$\mathcal{J}_{\text{CFB}}(P_{Y|X}, \beta) = I(X;Y) - \beta I(T;Y|S). \tag{6}$$

Proof. If we manipulate the expression of the CFB Lagrangian we can see how minimizing $\mathcal{L}_{\text{CFB}}(P_{Y|X}, \lambda)$ is equivalent to minimizing $\mathcal{J}_{\text{CFB}}(P_{Y|X}, \beta)$, where $\beta = \lambda + 1$. More specifically,

$$\begin{aligned}
\arg\inf_{P_{Y|X}\in\mathcal{P}} \left\{\mathcal{L}_{\text{CFB}}(P_{Y|X}, \lambda)\right\} &= \arg\inf_{P_{Y|X}\in\mathcal{P}} \left\{I(S;Y) + I(X;Y|S,T) - \lambda I(T;Y|S)\right\} && (17)\\
&= \arg\inf_{P_{Y|X}\in\mathcal{P}} \left\{I(X;Y) - I(T;Y|S) - \lambda I(T;Y|S)\right\} && (18)\\
&= \arg\inf_{P_{Y|X}\in\mathcal{P}} \left\{I(X;Y) - (\lambda+1) I(T;Y|S)\right\} && (19)\\
&= \arg\inf_{P_{Y|X}\in\mathcal{P}} \left\{\mathcal{J}_{\text{CFB}}(P_{Y|X}, \lambda+1)\right\}, && (20)
\end{aligned}$$

where $\mathcal{P}$ is the set of probability distributions over $\mathcal{Y}$ such that if $P_{Y|X} \in \mathcal{P}$ for all $X \in \mathcal{X}$, then the Markov chains $S \leftrightarrow X \rightarrow Y$ and $T \leftrightarrow X \rightarrow Y$ hold. □

A.2 Equivalence of the Lagrangians of the privacy funnel and the CPF

The privacy funnel is defined in a similar way to the CPF. It is an optimization problem that tries to design an encoding probability distribution $P_{Y|X}$ such that the representation Y keeps a certain level r' of information about the data of interest X, while minimizing the information it keeps about the private data S [2]. That is,

$$\arg\inf_{P_{Y|X}} \left\{I(S;Y)\right\} \text{ s.t. } I(X;Y) \geq r'. \tag{21}$$

Therefore, the Lagrangian of the privacy funnel problem is

$$\mathcal{L}_{\text{PF}}(P_{Y|X}, \alpha) = I(S;Y) - \alpha I(X;Y), \tag{22}$$

where $\alpha \in [0,1]$ is the Lagrange multiplier of $\mathcal{L}_{\text{PF}}(P_{Y|X}, \alpha)$.[5] This multiplier controls the trade-off between the information the representations keep about the private and the original data.

Proposition 3. *Minimizing $\mathcal{L}_{\text{CPF}}(P_{Y|X}, \lambda)$ is equivalent to minimizing $\mathcal{L}_{\text{PF}}(P_{Y|X}, \alpha)$, where $\alpha = \lambda/(\lambda+1)$.*

Proof. If we manipulate the expression of the CPF Lagrangian we can see how the minimizing $\mathcal{L}_{\text{CPF}}(P_{Y|X}, \lambda)$ is equivalent to minimizing $\mathcal{L}_{\text{PF}}(P_{Y|X}, \alpha)$, where $\alpha = \lambda/(\lambda+1)$. More specifically,

$$\begin{aligned}
\arg \inf_{P_{Y|X} \in \mathcal{P}} \left\{ \mathcal{L}_{\text{CPF}}(P_{Y|X}, \lambda) \right\} &= \arg \inf_{P_{Y|X} \in \mathcal{P}} \left\{ I(S;Y) - \lambda I(X;Y|S) \right\} && (23)\\
&= \arg \inf_{P_{Y|X} \in \mathcal{P}} \left\{ I(S;Y) - \lambda \left(I(X;Y) - I(S;Y) \right) \right\} && (24)\\
&= \arg \inf_{P_{Y|X} \in \mathcal{P}} \left\{ (\lambda+1) I(S;Y) - \lambda I(X;Y) \right\} && (25)\\
&= \arg \inf_{P_{Y|X} \in \mathcal{P}} \left\{ (\lambda+1) \left(I(S;Y) - \frac{\lambda}{\lambda+1} I(X;Y) \right) \right\} && (26)\\
&= \arg \inf_{P_{Y|X} \in \mathcal{P}} \left\{ \mathcal{L}_{\text{PF}} \left(P_{Y|X}, \frac{\lambda}{\lambda+1} \right) \right\}, && (27)
\end{aligned}$$

where $\mathcal{P}$ is the set of probability distributions over $\mathcal{Y}$ such that if $P_{Y|X} \in \mathcal{P}$ for all $X \in \mathcal{X}$, then the Markov chain $S \leftrightarrow X \rightarrow Y$ holds. □

We note how the relationship $\alpha = \lambda/(\lambda+1)$ maintains $\alpha \in [0,1)$ for $\lambda \geq 0$. This showcases how the CPF poses a more restrictive problem, in the sense that as long as $\lambda < \infty$ there are no solutions of the problem that filter private information arbitrarily.

B Modification of common algorithms to obtain private and/or fair representations

In this section of the supplementary material, we discuss the simple changes needed to common representation learning algorithms to implement our proposed variational approach. First, we show how common unsupervised learning algorithms can be modified to the variational approach to the CPF, thus generating private representations. Then, we show how common supervised learning algorithms can be modified to the variational approach to the CFB, thus generating fair representations.

Common unsupervised learning algorithms. The cost function of the β-VAE [16] and the VIB [24] (when the target variable is the identity of the samples) is

$$\mathcal{F}_{\text{uns}}(\theta, \phi, \eta) = \frac{1}{N} \sum_{i=1}^{N} D_{\text{KL}} \left(p_{Y|(X=x^{(i)},\theta)} || q_{Y|\theta} \right) - \eta^{-1} \mathbb{E}_{p_E} \left[\log \left(q_{X|(Y=f(x^{(i)},E),\phi)}(x^{(i)}) \right) \right], \tag{28}$$

where η is a parameter that controls the trade-off between the compression of the representations Y and their ability to reconstruct the original data X. Similarly, the VAE [15] cost function is $\mathcal{F}_{\text{uns}}(\theta, \phi, 1)$.

Comparison with the CPF. If we compare (28) with the cost function of the CPF $\tilde{\mathcal{J}}_{\text{CPF}}(\theta, \phi, \gamma)$, we observe that the only difference (provided that $\eta^{-1} = \gamma$) is the decoding density. In the CPF the decoding density of the original data X depends both on the representations Y and on the private data S, while in (28) it only depends on the representations Y. Therefore, the cost function $\mathcal{F}_{\text{uns}}(\theta, \phi, \eta)$

[5] If $\alpha = 1$, then $\mathcal{L}_{\text{PF}}(P_{Y|X}, 1) = -I(X;Y|S)$, for which optimal values of the encoding distribution $P_{Y|X}$ can filter private information arbitrarily. If $\alpha > 1$ this problem is even more pronounced. For $\alpha \leq 0$ trivial encoding distributions like a degenerate distribution $P_{Y|X}$ with density $p_{Y|X} = \delta(Y)$ are minimizers of the Lagrangian.

is recovered from the cost function of the CPF in the case that $q_{X|(S,Y,\phi)} = q_{X|(Y,\phi)}$. However, this is not desirable, since it means that the representations Y contain all the information from the private data S necessary to reconstruct X.

Modifications to obtain private representations. In these usupervised learning algorithms [15, 16, 24] the decoding (or generative) density is parametrized with neural networks, e.g., $q_{X|(Y,\phi)} = \text{Cat}(X; \rho_{\text{dec}}(Y;\phi))$ if X is discrete and $q_{X|(Y,\phi)} = \mathcal{N}(X; \mu_{\text{dec}}(Y;\phi), \sigma_{\text{dec}}(Y;\phi)^2 I_{d_{\text{dec}}})$ if X is continuous, where ρ_{dec}, μ_{dec}, and σ_{dec} are neural networks and d_{dec} is the dimensionality of X. In this work, the decoding density can also be parametrized with neural networks, e.g., $\text{Cat}(X; \rho'_{\text{dec}}(S,Y;\phi))$ if X is discrete and $q_{X|(Y,\phi)} = \mathcal{N}(X; \mu'_{\text{dec}}(S,Y;\phi), \sigma'_{\text{dec}}(S,Y;\phi)^2 I_{d_{\text{dec}}})$ if X is continuous, where ρ'_{dec}, μ'_{dec}, and σ'_{dec} are neural networks. Therefore, if the decoding density neural networks from [15, 16, 24] are modified so that they take the private data S as an input, then the resulting algorithm is the one proposed in this paper.

Common supervised learning algorithms. The cost function of the VIB [24] and the nonlinear IB [18] is

$$\mathcal{F}_{\text{sup}}(\theta,\phi,\eta) = \frac{1}{N}\sum_{i=1}^{N} D_{\text{KL}}\left(p_{Y|(X=x^{(i)},\theta)} || q_{Y|\theta}\right) - \eta^{-1}\mathbb{E}_{p_E}\left[\log\left(q_{T|(Y=f(x^{(i)},E),\phi)}(t^{(i)})\right)\right], \tag{29}$$

where η is a parameter that controls the trade-off between the compression of the representations Y and their ability to draw inferences about the task T.

Comparison with the CFB. Similarly to the comparison of the unsupervised learning algorithms and the CPF, we observe that (29) only differs with the cost function $\tilde{\mathcal{J}}_{\text{CFB}}(\theta,\phi,\beta)$ in the decoding density, i.e., the cost function $\mathcal{F}_{\text{sup}}(\theta,\phi,\eta)$ can be recovered from $\tilde{\mathcal{J}}_{\text{CFB}}(\theta,\phi,\beta)$ by setting $q_{T|(S,Y,\phi)} = q_{T|(Y,\phi)}$. The inference density of the task T only depends on the representations Y in [24, 18], while in our work it depends both on the representations Y and the sensitive data S. Hence, in these works the representations contain all the information from the sensitive data S necessary to draw inferences about the task T.

Modifications to obtain fair representations. The argument is analogous to the one for the modifications of unsupervised learning algorithms to obtain private representations. The only modification required in these supervised learning algorithms [24, 18] is to modify the decoding density neural networks to receive the sensitive data S as an input as well as the representations Y.

Invariants of the algorithms. In all these works [15, 16, 24, 18] and ours, the first (or the compression) term is usually calculated assuming that the encoder density is parametrized with neural networks, e.g., $p_{Y|(X,\theta)} = \mathcal{N}(Y; \mu_{\text{enc}}(X;\theta), \sigma_{\text{enc}}(X;\theta)^2 I_d)$, which allows the representations to be constructed using the reparametrization trick, e.g., $Y = \mu_{\text{enc}}(X;\theta) + \sigma_{\text{enc}}(X;\theta)E$, where $E \sim \mathcal{N}(0, I_d)$, d is the dimensionality of the representations, and I_d is the d-dimensional identity matrix. Then, the marginal density of the representations is set so that the Kullback-Leibler divergence has either a closed expression, a simple way to estimate it, or a simple upper bound, e.g., $q_{Y|\theta} = \mathcal{N}(Y; 0, I_d)$ or $q_{Y|\theta} = \frac{1}{N}\sum_{i=1}^{N} p_{Y|(X=x^{(i)},\theta)}$, where $x^{(i)}$ are the input data samples. Moreover, the loss function applied to the output of the decoding density and the optimization algorithm, e.g., stochastic gradient descent or Adam [45], can remain the same in these works and ours, too.

Remark 2. *The aforementioned modifications can also be introduced in other algorithms with cost functions with additional terms to $\mathcal{F}_{\text{uns}}$ and $\mathcal{F}_{\text{sup}}$. For example, adding a maximum-mean discrepancy (MMD) term on the represenation priors to avoid the information preference problem like in the InfoVAE [46]; adding an MMD term on the encoder densities to enforce privacy or fairness like in the VFAE [27]; or adding a total correlation penalty to the representation's marginal to enforce disentangled represenations like in the Factor-VAE, the β-TCVAE, or the FFVAE [47, 48, 36].*

C Non-convexity of the CPF and the CFB

In this section of the supplementary material, we show how both the CPF and the CFB as defined in (1) and (2) are non-convex optimization problems.

Lemma 1. *Let $X \in \mathcal{X}$, $S \in \mathcal{S}$, $Y \in \mathcal{Y}$, and $T \in \mathcal{T}$ be random variables. Then,*

1. *If the Markov chain $S \leftrightarrow X \rightarrow Y$ holds and the distributions of X and S are fixed, then $I(X;Y)$, $I(S;Y)$, and $I(X;Y|S)$ are convex functions with respect to the density $p_{Y|X}$.*
2. *If, additionally, the Markov chain $T \leftrightarrow X \rightarrow Y$ holds and the distributions of X, S, and T are fixed, then $I(X;Y|S,T)$ and $I(T;Y|S)$ are also convex functions with respect to the density $p_{Y|X}$.*

Proof. We start the proof leveraging [25, Theorem 2.7.4], which, in our setting, tells us that:

- $I(X;Y)$ is a convex function of $p_{Y|X}$ if p_X is fixed.
- $I(S;Y)$ is a convex function of $p_{Y|S}$ if p_S is fixed.
- $I(X;Y|S)$ is a convex function of $p_{Y|(X,S)}$ if $p_{X|S}$ is fixed.
- $I(X;Y|S,T)$ is a convex function of $p_{Y|(X,S,T)}$ if $p_{X|(S,T)}$ is fixed.
- $I(T;Y|S)$ is a convex function of $p_{Y|(S,T)}$ if $p_{T|S}$ is fixed.

Then, since $p_{Y|S} = \mathbb{E}_{p_{X|S}}[p_{Y|X}]$, $p_{Y|(X,S)} = \left(\frac{p_{X|S}}{p_X}\right) p_{Y|X}$, $p_{Y|(X,S,T)} = \left(\frac{p_{X|(S,T)}}{p_X}\right) p_{Y|X}$, and $p_{Y|(S,T)} = \mathbb{E}_{p_{X|(S,T)}}[p_{Y|X}]$ are non-negative weighted sums as defined in [23, 2.2.1], they preserve convexity. Hence, $I(S;Y)$, $I(X;Y|S)$, $I(X;Y|S,T)$, and $I(T;Y|S)$ are convex functions of $p_{Y|X}$, if p_S, $p_{X|S}$, $p_{X|(S,T)}$, and $p_{T|S}$ are fixed, respectively. □

Proposition 4. *Let us consider that the distributions of S and X are fixed and that the conditional distribution $P_{Y|X}$ has a density $p_{Y|X}$. Then, the CPF optimization problem is not convex.*

Proof. From Lemma 1 we know that $I(S;Y)$ and $I(X;Y|S)$ are convex functions with respect to $p_{Y|X}$ for fixed p_S and $p_{X|S}$. Hence, the constraint $I(X;Y|S) \geq r$ is concave. □

Proposition 5. *Let us consider that the distributions of S, T, and X are fixed and that the conditional distribution $P_{Y|X}$ has a density $p_{Y|X}$. Then, the CFB optimization problem is not convex.*

Proof. From Lemma 1 we know that $I(S;Y)$, $I(X;Y|S,T)$, and $I(T;Y|S)$ are convex functions with respect to $p_{Y|X}$ for fixed p_S, $p_{X|(S,T)}$, and $p_{T|S}$. Hence, the constraint $I(T;Y|S) \geq r$ is concave. □

D Details of the experiments

In this section of the supplementary material, we include an additional experiment on the COMPAS dataset [49] and describe the details of the experiments performed to validate the approach proposed in this paper.

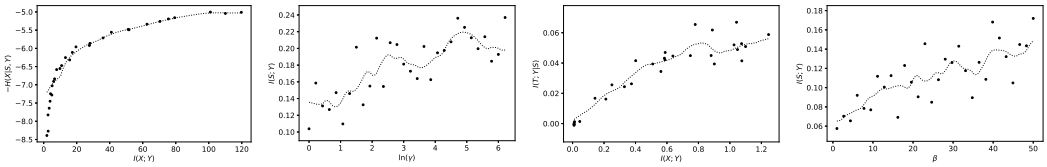
D.1 Results on the COMPAS dataset

COMPAS dataset. The ProPublica COMPAS dataset [49][6] contains $6,172$ samples of different attributes of criminal defendants in order to classify if they will recidivate within two years or not. These attributes include *gender*, *age*, or *race*. In both tasks, we followed the experimental set-up from Zhao et al. [10] and considered S to be a binary variable stating if the defendant is African American and X to be the rest of attributes. For the fairness task, we considered T to be the binary variable stating if the defendant recidivated or not. Since this dataset was not previously divided between training and test set, we randomly splitted the dataset with 70% of the samples $(4,320)$ for training and the rest $(1,852)$ for testing.

[6] Available at https://www.kaggle.com/danofer/compass.

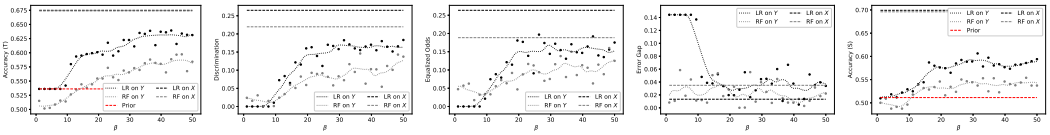
Similarly to the previous experiments, the proposed approach controls the trade-off between private and informative representations and between fair and accurate representations. In Figure 7 we see how the trade-off between the compression level $I(X;Y)$ and the informativeness of the representations independent of the private data $I(X;Y|S)$ and between the compression level $I(X;Y)$ and the predictability of the representations without the sensitive data $I(X;Y|S)$ is controlled by the private and the fair representations, respectively. Moreover, we can also see how the amount of information the representations keep about the private or the sensitive data is commanded by the Lagrange multipliers γ and β.

Furthermore, the Lagrange multiplier β also allows us to control the behavior of the accuracy, the error gap, and the discrimination for the COMPAS dataset (Figures 8a, 8d, and 8b). Moreover, in this scenario, as shown in Figures 8c and 8e, an increase of β also increased the equalized odds level and the accuracy on S of adversarial classifiers (even though they remained below their values obtained with the original data X for all the β tested). These results on the equalized odds, even though not generalizable since we have the counter-example of the Adult dataset, indicate that in some situations this quanitty can be controlled with our approach. More specifically, we believe this happens when we can guarantee that $I(S;Y;T)$ is non-negative as explained in Remark 4.



(a) (b) (c) (d)

Figure 7: Trade-off between (a) the private representations compression $I(X;Y)$ and the non-private information retained $I(X;Y|S)$ and (b) the Lagrange multiplier $\gamma \in [1, 500]$ and the private information $I(S;Y)$ kept by the representations. Since $I(X;Y|S) = H(X|S) - H(X|S,Y)$ and $H(X|S)$ does not depend on Y, the reported quantity is $-H(X|S,Y)$. Moreover, trade-off between (c) the private representations compression $I(X;Y)$ and the non-sensitive information retained about the task $I(T;Y|S)$ and (d) the Lagrange multiplier $\beta \in [1, 50]$ and the sensitive information kept by the representations. All quantities are obtained for the COMPAS dataset. The dots are the computed empirical values and the lines are the moving average of the 1D linear interpolations of such points.



(a) (b) (c) (d) (e)

Figure 8: Behavior of (a) the accuracy on T, (b) the discrimination, (c) the equalized odds gap, (d) the error gap, and (e) the accuracy on S of a logistic regression (LR, in black) and a random forest (RF, in gray) of the fair representations (dots and dotted line) and the original data (dashed line) learned with $\beta \in [1, 50]$ on the COMPAS dataset. The dashed line in red is the accuracy of a prior-based classifier.

D.2 Experimental details

Encoders. In all the experiments performed, we modeled the encoding density as an isotropic Gaussian distribution, i.e., $p_{Y|(X,\theta)} = \mathcal{N}(Y; \mu_{\text{enc}}(X;\theta), \sigma_\theta^2 I_2)$, so that $Y = \mu_{\text{enc}}(X;\theta) + \sigma_\theta E$, where $E \sim \mathcal{N}(0, I_2)$, μ_{enc} is a neural network, σ_θ is also optimized via gradient descent but is not calculated with X as an input, and where the representations have 2 dimensions. The neural networks in each experiment were:

- For the Adult dataset, μ_{enc} was a multi-layer perceptron with a single hidden layer with 100 units and ReLU activations.

Table 1: Convolutional neural network architectures employed for the Colored MNIST dataset. The network modules are the following: `Conv2D(cin,cout,ksize,stride,pin,pout)` and `ConvTrans2D(cin,cout,ksize,stride,pin,pout)` represent, respectively, a 2D convolution and transposed convolution, where `cin` is the number of input channels, `cout` is the number of output channels, `ksize` is the size of the filters, `stride` is the stride of the convolution, `pin` is the input padding of the convolution, and `pout` is the ouptut padding of the convolution; `MaxPool2D(ksize,stride,pin)` represents a max-pooling layer, where the variables mean the same than for the convolutional layers; `Linear(nu)` represents a linear layer, where `nu` are the output units; and `BatchNorm`, `ReLU6`, `Flatten`, `Unflatten`, and `Sigmoid` represent a batch normalization, ReLU6, flatten, unflatten and Sigmoid layers, respectively.

Name	Architecture
`CNN-enc-1`	`Conv2D(3,5,5,2,1,0) - BatchNorm - RelU6 - Conv2D(5,50,5,2,0,0) - BatchNorm - RelU6 - Flatten - Linear(100) - BatchNorm - ReLU - Linear(2)`
`CNN-enc-2`	`Conv2D(3,5,5,0,2,0) - BatchNorm - RelU6 - Conv2D(3,5,5,0,2,0) - BatchNorm - RelU6 - Conv2D(3,5,5,0,2,0)`
`CNN-dec-1`	`Linear(100) - BatchNorm - ReLU6 - Linear(1250) - Unflatten - BatchNorm - ReLU - ConvTrans2D(50,5,5,2,0,0) - BatchNorm - ReLU - ConvTrans2D(5,3,5,2,1,1) - Sigmoid`
`CNN-dec-2`	`Conv2D(3,5,5,0,2) - BatchNorm - RelU6 - Conv2D(5,50,5,0,2,0) - BatchNorm - RelU6 - Conv2D(5,50,5,0,2,0) - BatchNorm - RelU6 - Conv2D(5,50,5,0,2,0) - Sigmoid`
`CNN-mine`	`Conv2D(3,5,5,1,1,0) - MaxPool2D(5,2,2) - RelU6 - Conv2D(5,50,5,1,0,0) - MaxPool2D(5,2,2) - RelU6 - Flatten - Linear(100) - ReLU6 - Linear(50) - ReLU6 - Linear(1)`

- For the Colored MNIST dataset, μ_{enc} was the convolutional neural network `CNN-enc-1` for both the privacy and fairness experiments, and the convolutional neural network `CNN-enc-2` for the example from Figure 5. Both architectures are described in Table 1.
- For the COMPAS dataset, μ_{enc} was a multi-layer perceptron with a single hidden layer with 100 units and ReLU activations.

Moreover, the marginal density of the representations was modeled as an isotropic Gaussian of unit variance and zero mean; i.e., $q_{Y|\theta} = \mathcal{N}(Y; 0, I_2)$.

Decoders. In all the experiments performed for the fairness problem, the target task variable T was binary. Hence, we modeled the inference density with a Bernoulli distribution[7]; i.e., $q_{T|(S,X,\phi)} = \text{Bernoulli}(T; \rho_{\text{dec}}(S, Y; \phi))$, where ρ_{dec} is a neural network with a Sigmoid activation function in the output. In the privacy problem, if X was a collection of random variables $(X_1, X_2, ..., X_C)$, the generative density was modeled as the product of C categorical and isotropic Gaussians, depending if the variables were discrete or continuous. That is, $q_{X|(S,Y,\phi)} = \prod_{j=1}^{C} \text{Cat}(X_j; \rho_{\text{dec},j}(S, Y; \phi))^{\mathbb{I}[X_j=\text{Discrete}]} \mathcal{N}(X_j; \mu_{\text{dec},j}(S, Y; \phi), \sigma^2_{\phi,j})^{\mathbb{I}[X_j=\text{Continuous}]}$, where the continuous random variables are 1-dimensional. In practice, the densities were parametrized with a neural network with $K = \sum_{j=1}^{C} K_j^{\mathbb{I}[X_j=\text{Discrete}]} 1^{\mathbb{I}[X_j=\text{Continuous}]}$ output neurons, where K_j are the number of classes of the categorical variable X_j and each group of output neurons defines each multiplying density; either as the logits of the K_j classes or the parameter (mean) of Gaussian (the variances were also optimized via gradient descent but were not calculated with S nor Y as an input). If X was an image, the generative density was modeled as a product of $3C$ Bernoulli densities, where C is the number of pixels of the image and the 3 comes from the RGB channels. The neural networks in each experiments were:

- For the Adult dataset, the decoding neural network was a multi-layer perceptron with a single hidden layer with 100 units and ReLU activations. For the fairness task, the output was 1-dimensional with a Sigmoid activation function. For the privacy task, the output was 121-dimensional. The input of the network was a concatenation of Y and S.

[7]Note that the Bernoulli distribution is a categorical distribution with two possible outcomes.

Table 2: Hyperparameters employed to optimize the encoder and decoder networks of the experiments.

Dataset (task)	Epochs	Learning rate	Lagrange multiplier (β or γ)	Batch size
Adult (fairness)	150	10^{-3}	1-50 (logarithmically spaced)	1024
Adult (privacy)	150	10^{-3}	1-50 (logarithmically spaced)	1024
Colored MNIST (fairness)	250	10^{-3}	1-50 (logarithmically spaced)	1024
Colored MNIST (privacy)	500	10^{-3}	1-50 (logarithmically spaced)	1024
Colored MNIST (example)	500	10^{-3}	1	2048
COMPAS (fairness)	150	10^{-4}	1-50 (linearly spaced)	64
COMPAS (privacy)	250	10^{-4}	1-500 (logarithmically spaced)	64

Table 3: Hyperparameters employed to optimize the MINE networks of the experiments.

Dataset (task)	Iterations	Learning rate	Batch size
Adult (fairness)	$5 \cdot 10^4$	10^{-3}	2048
Adult (privacy)	$5 \cdot 10^4$	10^{-3}	2048
Colored MNIST (fairness)	$5 \cdot 10^4$	10^{-3}	2048
Colored MNIST (privacy)	$5 \cdot 10^4$	10^{-3}	2048
COMPAS (fairness)	$5 \cdot 10^4$	10^{-3}	463
COMPAS (privacy)	$5 \cdot 10^4$	10^{-3}	463

- For the Colored MNIST dataset and the fairness task, the decoding neural network was also a multi-layer perceptron with a single hidden layer with 100 units, ReLU activations, and a 1-dimensional output with a Sigmoid activation function. For the privacy task, the decoding neural network was the `CNN-dec-1` for the normal experiments and the `CNN-dec-2` for the example of Figure 5. The input linear layers took as an input a concatenation of Y and S and in the convolutional layers S was introduced as a bias.
- For the COMPAS dataset, the decoding neural network also was a multi-layer perceptron with a single hidden layer with 100 units and ReLU activations. For the fairness task, the output was 1-dimensional with a Sigmoid activation function. For the privacy task, the output was 19-dimensional. The input of the network was a concatenation of Y and S.

Hyperparameters. The hyperparameters employed in the experiments to train the encoder and decoder networks are displayed in Table 2, and the optimization algorithm used was Adam [45].

Preprocessing. The input data X for the Adult and the COMPAS dataset was normalized to have 0 mean and unit variance. The input data for the Colored MNIST dataset was scaled to the range $[0, 1]$.

Information measures. The mutual information $I(X;Y)$ and the conditional entropy $I(X;Y|S)$ and $I(T;Y|S)$ were calculated with the bounds from (7), (8), and (9), respectively. Since $H(X|S)$ was not directly obtainable, $-H(X|S,Y)$ was calculated and displayed instead. The mutual information $I(S;Y)$ was calculated using the mutual information neural estimator (MINE) with a moving average bias corrector with a an exponential rate of 0.1 [37]; the resulting information was averaged over the last 100 iterations. The neural networks employed were a 2-hidden layer multi-layer perceptron with 100 ReLU6 activation functions for all the datasets and tasks, except from the example on the Colored MNIST dataset, where the `CNN-mine` from Table 1 was used. In all tasks the input was a concatenation of S and Y, except from the example on the Colored MNIST dataset, where in all convolutional layers the private data S was added as a bias and in all linear layers S was concatenated to the input. The hyperparameters used to train the networks are displayed in Table 3.

Group fairness and utility indicators. The accuracy on T and S, the discrimination, and the error and equalized odds gaps were calculated using both the input data X and the generated representations Y. They were calculated with a Logistic regression (LR) classifier and a random forest (RF) classifier with the default settings from `scikit-learn` [50]. The prior displayed on the accuracy on T and S figures is the accuracy of a classifier that only infers the majority class of T and S, respectively, from the training dataset.

E Group fairness and utility indicators

In this section of the supplementary material, we define and put into perspective a series of metrics, employed in this article, that indicate the predicting and group fairness quality of a classifier.

Notation Let $X \in \mathcal{X}$, $S \in \mathcal{S}$, and $T \in \mathcal{T}$ be random variables denoting the input data, the sensitive data, and the target task data, respectively. Let also $\mathcal{X} \in \mathbb{R}^{d_X}$, $\mathcal{S} = \{0, 1\}$, and $\mathcal{T} = \{0, 1\}$. Let $w \in \mathcal{W}$, $w : \mathcal{X} \mapsto \mathcal{T}$ be a classifier; that is, w receives as an input an instance of the input data $x \in \mathcal{X}$ and outputs an inference about the target task value $t \in \mathcal{T}$ for that input data. Let us also consider the setting where we have a dataset that contains N samples of the random variables $D = \{(x^{(i)}, s^{(i)}, t^{(i)})\}_{i=1}^{N}$. Finally, let $\hat{P}$ denote the empirical probability distribution on the dataset D, $\hat{P}_{S=\sigma}$ the empirical probability distribution on the subset of the dataset D where $s^{(i)} = \sigma$, i.e., $\{(x, s, t) \in D : s = \sigma\}$, and $\hat{P}_{(S=\sigma,T=\tau)}$ the empirical probability distribution on the subset of the dataset D where $s^{(i)} = \sigma$ and $t^{(i)} = \tau$, i.e., $\{(x, s, t) \in D : s = \sigma \text{ and } t = \tau\}$.

A common metric to evaluate the performance (utility) of a classifier w on a dataset is its *accuracy*, which measures the fraction of correct classifications of the classifier on such a dataset.

Definition 1. *The* accuracy *of a classifier w on a dataset D is*

$$\text{Accuracy}(w, D) = \hat{P}(w(X) = T). \tag{30}$$

An ideally fair classifier w would maintain *demographic parity* (or *statistical parity*) and *accuracy parity*, which, respectively, mean that $w(X) \perp S$ (or, equivalently if w is deterministic, that $\hat{P}_{S=0}(w(X) = 1) = \hat{P}_{S=1}(w(X) = 1)$) and that $\hat{P}_{S=0}(w(X) \neq T) = \hat{P}_{S=1}(w(X) \neq T)$ [10]. In other words, if a classifier has demographic parity, it means that it gives a positive outcome with equal rate to the members of $S = 0$ and $S = 1$. However, demographic parity might damage the desired utility of the classifier [51], [12, Corollary 3.3]. Accuracy parity, on the contrary, allows the existance of perfect classifiers [10]. The metrics that assess the deviation of a classifier from demographic and accuracy parities are the *discrimination* or *demographic parity gap* [8, 10] and the *error gap* [10].

Definition 2. *The* discrimination *or* demographic parity gap *of a classifier w to the sensitive variable S on a dataset D is*

$$\text{Discrimination}(w, D) = \left| \hat{P}_{S=0}(w(X) = 1) - \hat{P}_{S=1}(w(X) = 1) \right|. \tag{31}$$

Definition 3. *The* error gap *of a classifier w with respect to the sensitive variable S on a dataset D is*

$$\text{Error gap}(w, D) = \left| \hat{P}_{S=0}(w(X) \neq T) - \hat{P}_{S=1}(w(X) \neq T) \right|. \tag{32}$$

Another advanced notion of fairness is that of *equalized odds* or *positive rate parity*, which means that $\hat{P}_{(S=0,T=\tau)}(w(X) = 1) = \hat{P}_{(S=1,T=\tau)}(w(X) = 1)$, for all $\tau \in \{0, 1\}$ or, equivalently, that $w(X) \perp S \mid T$ [51]. This notion of fairness requires that the true positive and false positive rates of the groups $S = 0$ and $S = 1$ are equal. The metric that assesses the deviation of a classifier from equalized odds is the *equalized odds gap* [10].

Definition 4. *The* equalized odds gap *of a classifier w with respect to the sensitive variable S on a dataset D is*

$$\text{Equalized odds gap}(w, D) = \max_{\tau \in \{0,1\}} \left| \hat{P}_{(S=0,T=\tau)}(w(X) = 1) - \hat{P}_{(S=1,T=\tau)}(w(X) = 1) \right|. \tag{33}$$

Remark 3. *In the particular case of learning fair representations, the classifier $w : \mathcal{X} \mapsto \mathcal{T}$ consists of two stages: an encoder $w_{\text{enc}} : \mathcal{X} \mapsto \mathcal{Y}$ and a decoder $w_{\text{dec}} : \mathcal{Y} \mapsto \mathcal{T}$, where the intermediate variable $Y = w_{\text{enc}}(X)$ is the fair representation of the data. Therefore:*

1. *Minimizing $I(S; Y)$ encourages demographic parity, since*

$$I(S; Y) = 0 \iff Y \perp S \implies w(X) \perp S. \tag{34}$$

2. *Minimizing $I(S; Y|T)$ encourages equalized odds, since*

$$I(S; Y|T) = 0 \iff Y \perp S \mid T \implies w(X) \perp S \mid T. \tag{35}$$

Remark 4. *Based on Remark 3, we note that the variational approach to the CFB and the CPF for generating private and/or fair representations encourages demographic parity, since the minimization of the Lagrangians of such problems, $\mathcal{L}_{\text{CFB}}$ and $\mathcal{L}_{\text{CPF}}$, indeed minimizes $I(S;Y)$.*

However, we cannot say the same for equalized odds. Even though $I(S;Y) = I(S;Y|T) + I(S;Y;T)$, since $I(S;Y;T)$ can be negative [22], then $I(S;Y)$ is not necessarily greater than $I(S;Y|T)$ and thus there is no guarantee that minimizing $I(S;Y)$ will minimize $I(S;Y|T)$ as well.