Learning to Conceal: A Method for Preserving Privacy and Avoiding Prejudice in Images

Avigail Stekel, Moshe Hanukoglu, Aviv Rovshitz, Nissan Goldberg, and Amos Azaria
Computer Science Department and Data Science Center
Ariel University, Israel

***Abstract*—We introduce a learning model able to conceal personal information (e.g. gender, age, ethnicity, etc.) from an image while maintaining any additional information present in the image (e.g. smile, hair-style, brightness). Our trained model is not provided the information that it is concealing, and does not try learning it either. Namely, we created a variational autoencoder (VAE) model that is trained on a dataset including labels of the information one would like to conceal (e.g. gender, ethnicity, age). These labels are directly added to the VAE's sampled latent vector. Due to the limited number of neurons in the latent vector and its appended noise, the VAE avoids learning any relation between the given images and the given labels, as those are given directly. Therefore, the encoded image lacks any of the information one wishes to conceal. The encoding may be decoded back into an image according to any provided properties (e.g. a 40-year old woman).**

Our method successfully conceals the private information; a convolutional neural network trained on the concealed images cannot restore the original private information. In contrast to the private information, a user study shows that the remaining properties of the original image carry-on to the concealed image.

The proposed architecture can be used as a mean for privacy preserving and can serve as an input to systems, which will become unbiased and not suffer from prejudice. We believe that privacy and discrimination are two of the most important aspects in which the community should try and develop methods to prevent misuse of technological advances.

I. Introduction

There are many implications of user privacy with respect to user data; foremost is the fear of exposing personal information over a social network. As indicated by Almadhoun et al. [2] 75.8% of the respondents do not believe that they would feel totally safe when providing sensitive information about themselves over the social networks. Indeed, several network attacks exist that allow strangers to extract personal information from a victim [21]. Therefore, any personal data uploaded to the internet might be exposed by a third party who should not be permitted to view it. Data anonymization methods could support user privacy.

In addition, user-privacy also relates to research communities. Rothstein states that the current regulatory frameworks of the Common Rule and Privacy Rule emphasize privacy interests, but they overlook the privacy interests of individuals whose health information and biological specimens are used in research without their knowledge, consent, or authorization [14]. Methods for data anonymization would allow faster and more effective research, including life-saving and medical research, as people are more likely to be willing to share truly anonymized data.

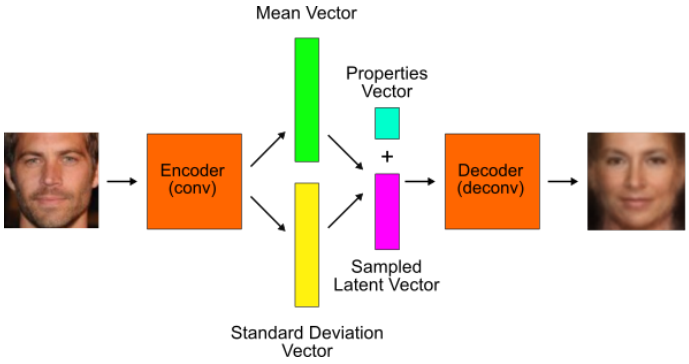

Figure 1. General architecture of The Blind Autonencoder for Fairness and Objectiveness (BAFO).

Another important problem addressed in this paper is the issue of discrimination. According to an Analysis commissioned by the British daily newspaper The Guardian, information from the London Mayor's Office for Policing and Crime, shows that while black people compose only 15.6% of London's population (and white people compose 59.8% of them), in 2018, 43% of searches were of black people, while only 35.5% were of white people [16]. Furthermore, searches of black people were less likely to detect crime than those conducted on white people. As obtained from the data, while 21% of searches of white people led to an arrest, only 16% of searchs on black people led to an arrest. In addition, it seems that "disproportionality has increased", with the likelihood of black people being stopped being 4.3 times higher than white people in 2018, compared with 2.6 times more likely in 2014 [17]. Unfortunately, it has been shown that computer programs and machine learning algorithms suffer from prejudice and gender bias as well [9], [4], [18].

To overcome the issues of privacy and prejudice in images, we introduce the Blind Autonencoder For Fairness and Objectiveness (BAFO), a novel deep learning architecture that is based on a variation autoencoder (VAE) [11]. Namely, BAFO is trained on a dataset including labels of the information one would like to conceal (e.g. gender, ethnicity, age). These labels are directly added to the VAE's sampled latent vector (see Figure 1). Due to the limited number of neurons in the latent vector and its appended noise, the VAE avoids learning

any relation between the given images and the given labels, as those are given directly. Therefore, the encoded image lacks any of the information one wishes to conceal; BAFO is practically blind to all this information. The encoding may be decoded back into an image according to any provided properties (e.g. a 40-year old woman).

It might seem unintuitive that during training BAFO is given the information that it should later conceal. However, BAFO's behavior might be similar to a child that is asked to use a calculator from the very beginning of preschool. The child might focus on acquiring complex mathematical skills but is not likely to know how to add or multiply numbers by herself, because she learns to trust the calculator instead. Furthermore, if that child will later use a different calculator that uses different functions, the computation results will become different. Another similar example is the autopilot, which may cause pilots to not be able to fly an airplane themselves, because they learn to trust the autopilot [5].

Using BAFO, security offices may monitor concealed surveillance footage, that is, surveillance footage in which all information required to be concealed (e.g. gender, race) is not present. It is important to note that this information is not only removed, but explicitly not learned by BAFO. Furthermore, such footage may be fed into an artificial intelligence system that will detect suspicious act, but will be totally unbiased, as the footage will not include any of the concealed information, or even any information that will allow the system to deduct the concealed information from.

The general idea behind BAFO is to explicitly provide the model, during its learning process, with the information that it should conceal during the inference phase, in which this information will not be provided. We believe that this idea is not limited to images and videos, but is more general and can easily adapted to be used for concealing one's voice, and text. Concealing voice may be used by the press or by court when the identity of the speaker needs to remain unknown. Concealed text may be used when applying for education or job openings. A candidate may fear to fall for prejudice, and may therefore wish to hide her gender or race from the curriculum vitae (CV), in a way that this information cannot be deduced. Text concealing can be used also for robustly annonymizing data. This is because BAFO would remove not only any explicit identifiers (e.g. name and country of birth), but also from any implicit ones (e.g. specific expressions used by some group of people). In addition, users knowing that their data is totally annonymized (either by a company or even by running a system simalar to BAFO themselves) are more likely to share their data. Data sharing is especially important in any medical related system, which may literally save people's life.

It is very important to note that the idea behind BAFO is very different than other tools that may be used to convert one type of image to another (e.g. showing an older version of one-self) [1]. Such technologies apply a smart filter that converts some type of image to another, this filter is applied regardless of the original image. Therefore, in such systems if a feminine filter is applied on a woman, she would seem even more feminine. Unlike with BAFO, if trained to remove gender, in which all gender related aspects of any image are totally removed, and are later explicitly added in order to produce a new image.

We show that BAFO successfully conceals the private information; a convolutional neural network trained on the concealed images cannot restore the original private information. We further show that the concealed image does match the additional information provided. That is, for example, a male image concealed as a female, is usually classified as a female. A user study shows that the remaining properties of the original image carry-on to the concealed image; in 90% of the images, the participants could identify the original image that was concealed by BAFO to a given concealed image.

II. Related work

Zemel et al. [19] introduced a measure, $yDiscrim$, of discrimination for a classification problem. $yDiscrim$ is the difference between the ratio of the positive predictions in a specific set and the ratio of the positive predictions in the rest of the examples in the data. Formally:

$$yDiscrim = \left| \frac{\sum_{n:s_n=1} \hat{y}_n}{\sum_{n:s_n=1} 1} - \frac{\sum_{n:s_n=0} \hat{y}_n}{\sum_{n:s_n=0} 1} \right| \quad (1)$$

where S is a binary variable representing whether a given individual is a member of a specific set, and $\hat{y}$ is the prediction for y. We note that this measure is only applicable to a scenario in which the overall goal of a system is to classify individuals to different classes. Furthermore, this measure is not relevant for cases in which different groups must have different ratios. For example, a medical system predicting whether a subject is suffering from a specific illness. While in such cases we might wish to conceal private information, the system should still provide the correct ratio for each group. However, in our work, the overall goal of the system is to produce an output that conceals some information while remaining interpretable to humans.

Several fair machine learning methods have been developed recently including some based on deep learning. Louizo et al. [10] develop a classifier that is based on a variational autoencoder, VFAE. This classifier obtains a sensitive variable along with additional insensitive data and outputs a prediction of a class for each input. The authors apply VFAE to four datasets. For example, one of the datasets the authors apply VFAE to is the Adult income dataset. In this domain, VFAE obtains the age of an individual along with 14 additional attributes, and outputs a prediction as to whether the individual has an annual income of over $50{,}000$ or not. This work differs from ours in several aspects. First, VFAE does not output any intermediate value that is human interpretable. Secondly, VFAE is trained for a specific classification task. Finally, VFAE requires the sensitive information at inference, while BAFO uses the sensitive information only at training, and can therefore conceal this sensitive information without obtaining it. We also note that VFAE is limited to a binary classification task, that is, a task with only two classes. This limitation was

later relaxed by [15] who improved their method and extended it to a classifier with multiple classes.

Some previous works have developed different machine learning architectures with an attempted to model the age progress in images [12], [13], [20], [3]. Zhang et al. [20] developed a Conditional Adversarial AutoEncoder (CAAE) model that studies the facial features, and important parameters that appear in each age segment. When given an image it may be converted to a new image by progression or regression of the current age. In the first stage the face is mapped to an invisible vector, encoded by conventional coding. The hidden layer retains the characteristics of the facial features in a "personal" manner. There are two networks that improve each other and finally compare the result obtained with the real one. While Zhang et al. is, to the best of our knowledge, the only work that does not require to be given the age of a given image in order to output a specific age. Their work is still very different than ours, as for training, their model requires many images of the same person at different ages. CAAE uses the dataset to learn how age progresses / regresses and creates an age manifold that can then be used to create a new image.

In recent years there have been a number of popular approaches for creating artificial images: Generative Adversarial Network (GAN) [7] is a Generative model for creating new information such as creating fake high quality images. GANs are trained to output images that look real, and are therefore sharp and have high contrast. GANs include a discriminator and a generator; both components compete with each-other. The discriminator identifies whether each image is original or has been created by the generator, while the generator's goal is to create images that will seem real to the discriminator. That is, the generator tries to deceive the discriminator into thinking that the the images created by the generator are original. Baek et al. [3] created a face editing tool that is based on GANs. However, GANs are not appropriate for our goal, since we do not intend to produce an image that looks as real as possible, but to preserve the original image while concealing the intended properties.

Another deep learning based approach for generating images is the Variational AutoEncoder (VAE) [6]. Similar to a standard AutoEncoder, a VAE architecture includes a bottleneck and, during training, it tries to restore a given image. The representation in the bottleneck is in fact a compressed representation of the given image. However, a VAE learns two vectors, a mean vector and a standard deviation vector, the latent vector is sampled using the mean and standard deviation vectors. VAEs have been used for denoising and generating images that are similar to the training-set [8]. In this paper we propose a novel concept in which the VAE that is trained on the given data, but the labels are appended directly to the latent vector (see Figre 1). The VAE, therefore, does not attempt to learn any information that is directly provided, and becomes blind to these properties. While not being the primary intention of our VAE architecture, we note that it can also be used to generate images with specific attributes (e.g. only images of a 40-year-old female).

III. Dataset

We use the UTKFace dataset [20], which contains approximately 23,000 headshot images. The images in the dataset are labeled with the age (ages range from 0 to 116), gender (male and female) and ethnic origin which is divided into five types. The dataset was split into 85% training-set and 15% test-set. All our software is available at: https://github.com/avigailst/Learning_to_Conceal.

IV. Method

In this section we introduce the Blind Autonencoder For Fairness and Objectiveness (BAFO).

The general architecture of the model is depicted in Figure 2. In the training phase, a labeled image is inserted as an input to BAFO. In the UTKFace dataset the images size is 56x56x3 (RGB), and the images are tagged by age, gender, and ethnicity. We note that the labels include the information that we intend to later conceal. The image is then compressed by the encoder into two vectors representing the parameters that describe the image, one vector for the mean and the other for the standard deviation. After sampling the latent vector from the mean and the standard deviation, it is concatenated to the the information BAFO is concealing, the age, the gender and the ethnicity. The latent vector is then decoded back to an image with a size similar to the input image.

The motivation behind this architecture is that, since the information to be concealed is provided without noise, the system will be able to devote all of its efforts to learning only the additional parameters that affect the image, and not the information that is explicitly provided.

1) Latent Vector Size: We consider two different sizes for the latent vector, 48 and 100. In order to evaluate the performance of BAFO and determine which vector size to use, we concealed the test data and decoded it into females and males in 5 age groups: 1-year-old, 20-year-old, 40-year-old, 60-year-old and 80-year-old. In order to evaluate the decoded concealed images with respect to the age, we developed an age classifier. We note that the age classifier itself had a root mean squared error (RMSE) of 6.75. The root mean squared error (RMSE) and mean absolute error (MAE) of the decoded concealed images appear in Table I. As depicted by the table, BAFO with a latent vector of size 48 seems to slightly outperform BAFO with a latent vector of size 100.

V. Results

Figure 3 presents 7 randomly picked images from the test-set, and the output of BAFO when concealed as a 40-year old woman, using 48 and 100 length latent vectors. The input images appear in the first row; in the second row are the images concealed by BAFO when using a 48 length latent vector; and in the third row are the images concealed by BAFO when using a 100 length latent vector. Note that the facial features are smoother and delicate in the cheek and nose areas, as well as the thinner eyebrows in pictures 1,4 and 6, as shown in Figure 3. Note the way the system conceals only the age and gender of the given image, but preserves the smile,

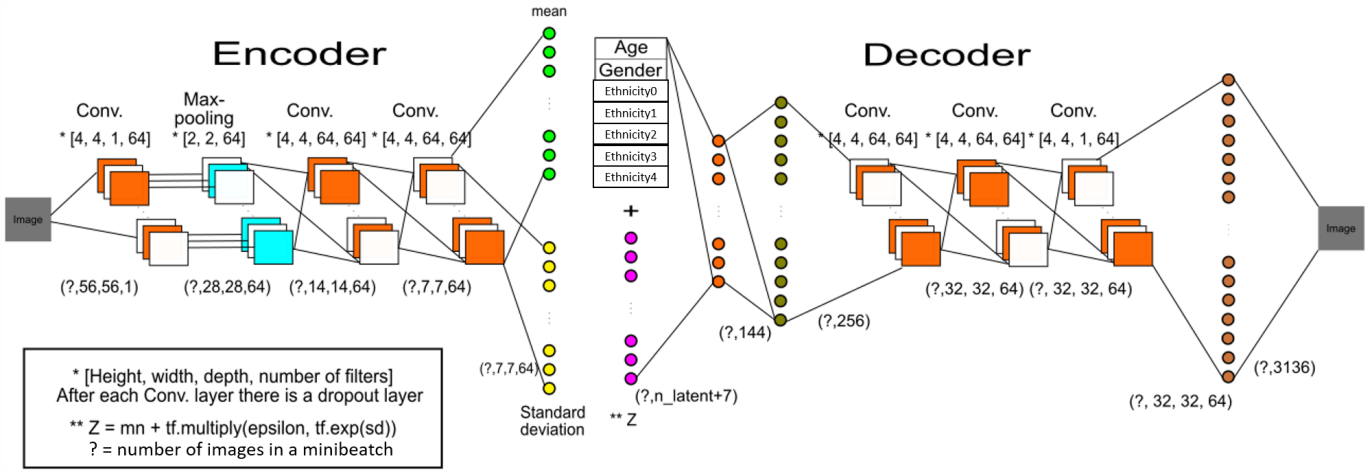

Figure 2. The Architecture of the VAE.

Target age	RMSE 48-cells	RMSE 100-cells	MAE 48-cells	MAE 100-cells
1	**12.62**	13.34	10.60	11.42
20	9.37	**9.00**	7.94	**7.70**
40	**7.89**	7.97	**6.56**	6.65
60	**18.03**	18.71	**16.45**	17.45
80	**27.41**	28.30	**26.03**	27.26
Average	**15.06**	15.47	**13.52**	14.10

Table I
THE ROOT MEAN SQUARED ERROR (RMSE) AND MEAN ABSOLUTE ERROR (MAE) OF CONCEALED IMAGES DECODED AS 1 YEAR OLDS, 20, 40 60 AND 80 YEAR OLDS (LOWER IS BETTER).

including the presence of teach, hair-style, light conditions, brightness of the images and the background.

Figure 4 presents the same images when concealed as a 40-year-old male. The male's facial features became more coarse in the center of the face, accompanied by wrinkles at the sides of her eyes and mouth, as can be seen in the second, fifth and seventh images. Note the smile, the length of the eyebrows and the structure of the nose are preserved in the second row after the process is completed.

In Figure 5, we used a single image as the input and concealed it as a male and a female with ages varying from a one-year-old to an 80-year-old person. As seen in the figure, BAFO can conceal any image at any range of ages and both as a male and a female. This result may resemble common image editor tools (especially when making an image look older or younger) [1], which are mainly used for entertainment. However it is very important to note that BAFO is very different than those tools as it is not trained to convert images from one demographic group to another (or making an image older or younger), but is totally blind to the demographic group. Other image editor tools are either directly trained to convert from one demographic group to another, or to add (and remove) specific features (such as aging features, feminine features, smile features etc.). Very importantly, BAFO is not provided the age, gender or ethnicity of the original image, nor does it learn it itself, as BAFO is not required to know whether the image should be converted to a younger image or an older image. Therefore, BAFO is truly unbiased and the demographic group is explicitly *added* to the image encoding only in order to produce a meaningful image.

VI. EVALUATION

There are three dimensions for the evaluation of BAFO. The first is an evaluation of the quality of output with respect to the demographic information provided. That is, if a male image was concealed as a female, the output image should be classified as a female. The second is the evaluation of BAFO's ability to conceal information. That is, determining to what extent the concealed information can be retrieved from the generated image. The third dimension for evaluating BAFO is its ability to construct an image that is similar to all properties of the original image, excluding the concealed information. The former two dimensions were evaluated by using a convolutional neural network, while the later dimension was evaluated by human judges.

A. Evaluation with a Convolutional Neural Network

To evaluate the first and the second dimensions, we developed a classifier with two convolution layers and two fully connected neural network layers. The classifier was trained and tested on different datasets according to its evaluation goal. We focus on the evaluation of the gender property of the image. For that end we identified four different data groups (as shown in Figure 6):

The images of *males* from the original dataset (group A).
The images of *females* from the original dataset (group B).
The images from groups A and B that were concealed as *males* by BAFO (group C).

Figure 3. Images concealed by BAFO as a woman 40-year-old. In the first row are the input images; in the second row are the images concealed by BAFO when using a 48 length latent vector; and in the third row are the images concealed by BAFO when using a 100 length latent vector. Note the way the system conceals only the age and gender of the given image, but preserves the smile, including the presence of teach, hair-style, light conditions, brightness of the images and the background.

Figure 4. Concealing images as a 40-year-old male. In the first row are the input images and in the second row are the images concealed by BAFO when using a 48 length latent vector.

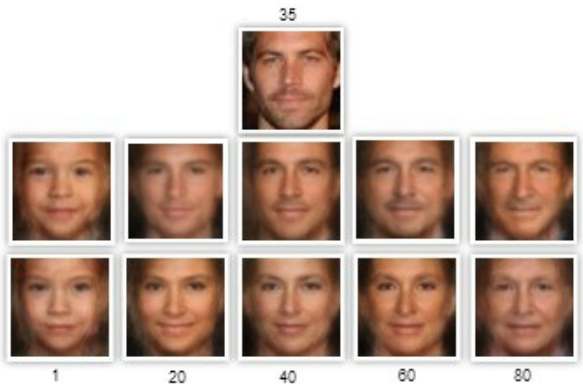

Figure 5. Concealing a single image as a man and a woman at different ages. This may resemble common image editor tools, however BAFO is very different than those tools as it is not trained to convert images from one demographic group to another, but is totally blind to the demographic group. The demographic group is explicitly *added* to the image encoding in order to produce a meaningful image.

The images from groups A and B that concealed as *females* by BAFO (group D).

All data groups were split into 90% training data and 10% test data. The classifier was first trained on the training data in groups A and B, with the label being the gender of the person in the images. When tested on the test data (of groups A and B) the classifier obtained an accuracy of 88:3%. A similar accuracy (86:5%) was obtained when tested on the images in groups C and D. Where in group C the classifier had to predict the value for male and in group D, female. This implies that BAFO's output matches the required gender. That is, BAFO succeeded in the evaluation of the first dimension.

In order to evaluate the second dimension, the classifier was then trained separately on groups C and D, with an attempt to recover the concealed information. That is, the images in groups C and D were labeled according to the gender in the original images. The classifier was tested on the test-set from groups C and D. In this evaluation lower accuracy means that the classifier could not recover the concealed information. That is, lower accuracy implies that BAFO had succeeded

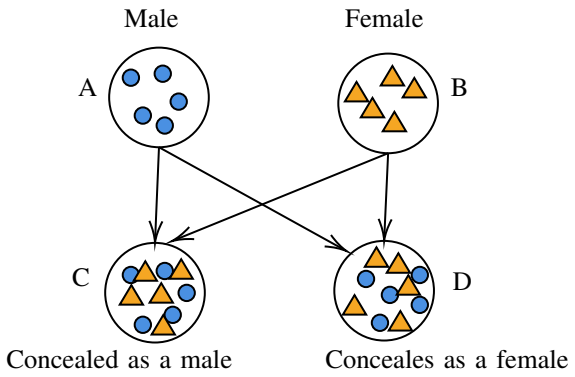

Figure 6. This figure illustrates the datasets that where used to evaluate BAFO using a convolutional neural network. The blue circles represent the male images and orange triangles represent female images. Group A and group B represent the original images, A for the male images and B for the female images. Groups C and D represent the images of A and B that were concealed as males in group C and as females in group D.

in concealing the private information. Indeed, the accuracy measured was very low (less than 50%, which could be obtained by a random classifier). All results are shown in II.

	Training	Test
$A+B \rightarrow A+B$	98%	88.3%
$A+B \rightarrow C+D$	$--$	86.5%
$C \rightarrow C$	86.5%	43.9%
$D \rightarrow D$	88.5%	47.7%

Table II

THE RESULTS OF THE CLASSIFIER THAT WAS DEVELOPED FOR EVALUATING BAFO BY THE FIRST AND THE SECOND DIMENSIONS. WHEN THE CLASSIFIER WAS TRAINED ON A+B AND TESTED ON THE TEST-SET OF A+B, ITS ACCURACY WAS 88.3%, WHICH IS VERY SIMILAR TO ITS ACCURACY WHEN TESTED ON GROUPS C+D. THIS IMPLIES THAT BAFO'S OUTPUTS MATCH THE REQUIRED GENDER. FURTHERMORE, THE CLASSIFIER DID NOT SUCCEED IN IDENTIFYING WHETHER AN IMAGE FROM GROUP C WAS CONCEALED FROM A MALE IMAGE OR A FEMALE IMAGE (ACCURACY 43.9%). A SIMILAR RESULT WAS OBTAINED IN GROUP D. THIS IMPLIES THAT BAFO CONCEALS THE GENDER PROPERTY VERY WELL.

B. Evaluation with Human Judges

In order to evaluate the third dimension we conducted a survey. The participants were asked ten questions: In the first set of five questions the participants were shown an original image from the dataset and two images that were transformed, by BAFO, to a 40-year-old woman. The participants were asked to choose the image which they believed was transformed from the original given image. The structure of the second set of five questions was reversed; we provided one transformed image and two original images and the participants were asked to pick an original image. A screenshot of the survey is presented in Figure 7. There were 127 participants aged between 17 and 72, 49 males and 80 females. The results are presented in the Table III. The participants answered correctly on 85.5% questions when the images that were in the questions were of the same ethnic group and 92.4% on questions with images not in the same ethnic group. The results show that BAFO's outputs preserve their properties, as humans successfully identify the original image even when the ethnicity property is the same.

	No. of queries	Success rate
From the same ethnic	315	85.8%
From a different ethnic	916	92.4%
Identifying the original image	670	88.8%
Identifying the concealed image	597	92.4%
Total	1267	90.5%

Table III

THE RESULTS FROM THE HUMAN SURVEY. THE RESULTS SHOW THAT BAFO'S OUTPUT PRESERVES THE IMAGE'S PROPERTIES AS HUMANS SUCCESSFULLY IDENTIFY THE ORIGINAL IMAGE (AND VICE-VERSA).

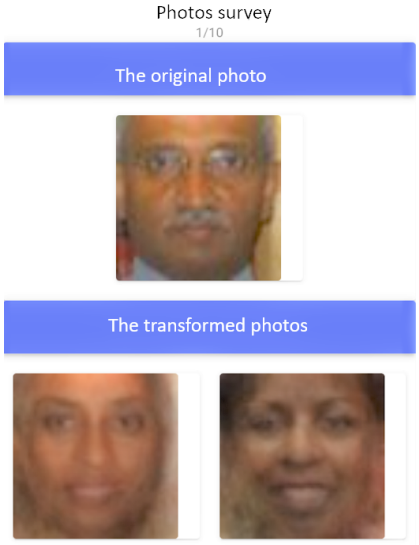

Figure 7. A screenshot from the published survey. In the top section there is an original image from the dataset. In the bottom section there are two images that were conceal to a 40-years old woman. Only one image in the bottom section was conceal from the top image. The survey participant was asked to click on the image that she believes was concealed from the top section

VII. DISCUSSION

As depicted by Table I and Figure 3, when modifying the latent vector size, there seems to be a trade-off between the image quality and the concealing performance. That is, with a large latent vector, the image decoded image quality is slightly better, but the image is slightly not concealed as well. This is expected, as the larger the latent vector, the more information BAFO can be stored in it, and the less does it need to rely on the additional information. On the other hand, the larger

the latent vector, the more features may BAFO store in it and the higher the image quality. It is yet to be determine what the optimal latent vector size is, this may depend not only at the application, but also at the amount of labeled data available. Another approach is to increase the latent vector size, but also to modify the KL-divergence formula, so that the standard deviation will receive higher penalty rates, and therefore the data obtained from the mean-vector will be very noisy. This will further encourage BAFO to rely on the provided information, as much as possible.

In order to decode the concealed images, so that it is understandable for humans, BAFO needs to be given some demographic information (or any other information related to the concealed information). However, it would be preferable if BAFO could present the information using a neutral representation. For example, by using an image that is not gender associated. As a first approach to achieve such an image, we simply provided a value of 0.5 (the mean of the values for female and for male) as the gender value to BAFO (See Figure 8). However, as can be seen in the figure this approach did not perform that well; this is not surprising, since no image with the value of 0.5 was given to BAFO during the training phase.

VIII. Conclusions & Future Work

In this paper we introduce BAFO, an image concealer that receives as input an image and conceals unwanted properties (such as gender, ethnic origin, and age). The concealed images may be viewed by humans in a way that would remove any prejudice related to the concealed properties. BAFO may be embedded in smart glasses for security officers or other law enforcers, such that some properties of people who they interact with are concealed. These images can serve as input to another machine learning system, which, due to the input it receives from BAFO, will be unbiased. BAFO may also be used as a mean for privacy preserving by social network users. A user may conceal user private information in images (e.g. age, gender, ethnic origin) before she uploads them to a social network. Users who are familiar with that user and know the private information, will be able to decode the image according to the private information, and will view an image that is very similar to the original image that was uploaded. Other people will see the images, but will not be able to extract the private information concealed in these images.

Extending the architecture described in this paper to concealing one's identity is straight-forward. Such concealed images should preserve privacy and therefore could be used by researchers in different fields. Future work will also include the extension of BAFO's architecture to video, voice and text. Such extensions may have major implications on privacy preserving and unbiased systems, such as an *unbiased* surveillance camera with automatic security threat detection.

IX. Acknowledgment

This research was supported in part by the Ministry of Science, Technology & Space, Israel.

References

[1] Zahid Akhtar, Dipankar Dasgupta, and Bonny Banerjee. Face authenticity: An overview of face manipulation generation, detection and recognition. In *Nutan College of Engineering & Research, International Conference on Communication and Information Processing (ICCIP)*, 2019.

[2] Nour Mohammed Almadhoun, P Dhanapal Durai Dominic, and Lai Fong Woon. Perceived security, privacy, and trust concerns within social networking sites: The role of information sharing and relationships development in the malaysian higher education institutions' marketing. In *2011 IEEE International Conference on Control System, Computing and Engineering*, pages 426–431. IEEE, 2011.

[3] Kyungjune Baek, Duhyeon Bang, and Hyunjung Shim. Editable generative adversarial networks: Generating and editing faces simultaneously. In *Asian Conference on Computer Vision*, pages 39–55. Springer, 2018.

[4] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[5] Nicholas Carr. *The glass cage: Automation and us*. WW Norton & Company, 2014.

[6] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[8] Huaibo Huang, Ran He, Zhenan Sun, Tieniu Tan, et al. Introvae: Introspective variational autoencoders for photographic image synthesis. In *Advances in Neural Information Processing Systems*, pages 52–63, 2018.

[9] Susan Leavy. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, pages 14–16. ACM, 2018.

[10] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[11] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems*, pages 2352–2360, 2016.

[12] Narayanan Ramanathan and Rama Chellappa. Modeling age progression in young faces. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 387–394. IEEE, 2006.

[13] Narayanan Ramanathan and Rama Chellappa. Modeling shape and textural variations in aging faces. In *2008 8th IEEE International Conference on Automatic Face*