European Lighthouse of AI for Sustainability

Deliverable number 3.1
Date: 28 August 2024

First release of AI tools for Trustworthy AI

Project Title	ELIAS - European Lighthouse of AI for Sustainability
Contract No.	101120237
Start of Project	1 September 2023
Duration	48 months

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101120237.

Funded by
the European Union

Deliverable title	First release of AI tools for Trustworthy AI
Deliverable number	D3.1
Deliverable version	2.0
Previous version(s)	1.0
Contractual date of delivery	August 31, 2024
Actual date of delivery	August 28, 2024
Deliverable filename	ELIAS_D3.1.pdf
Nature of deliverable	Report
Dissemination level	Public
Number of pages	93
Work Package	WP3
Task(s)	T3.2, T3.3, T3.4
Parner responsible	FBK
Author(s)	
Editor	
Project Officer	Evangelia Markidou

Abstract	This deliverable presents the results at M12 for tasks T3.2 (Fairness and Counterfactual Interventions), T3.3 (Cognition-aware hybrid decision-making systems) and T3.4 (Privacy-Preserving Machine Learning).
Keywords	fairness and counterfactual interventions, trusworthy artificial intelligence, machine biases and hallucinations, cognitive biases, privacy-preserving machine learning, multimodal foundation models, computer vision, machine learning

Copyright

Contributors

NAME	ORGANIZATION
Bruno Lepri	FBK
Massimiliano Luca	FBK
Simone Centellegher	FBK
Veronica Lachi	FBK
Giovanni De Toni	FBK,UNITN
Rita Cucchiara	UMORE
Lorenzo Baraldi	UMORE
Enver Sangineto	UMORE
Federico Cocchi	UMORE
Tobia Poppi	UMORE
Nicu Sebe	UNITN
Roberto Zamparelli	UNITN
Raffaella Bernardi	UNITN
Andrea Passerini	UNITN
Nuria Oliver	ALC
Aditya Gulati	ALC,FBK
Gergely Nemeth	ALC
Adrian Arnaiz-Rodriguez	ALC
Joao Pitacosta	JSI
Inna Novalija	JSI
Emmanouil Krasanakis	CERTH

Peer Reviews

NAME	ORGANIZATION
Andrea Passerini	UNITN
Ulf Lüder	University of Tubingen

Revision History

Version	Date	Reviewer	Modifications
1.0	02/08/2024	FBK	Initial version, Table of contents and sections
2.0	28/08/2024	FBK	Revised version

Table of Abbreviations and Acronyms

Abbreviation	Meaning
AI	Artificial Intelligence
API	Application Programming Interface
AR	Algorithmic Recourse
AVG	Average
CLIP	Contrastive Language-Image Pretraining
CNN	Convolutional Neural Network
CSCF	Consequence-aware Sequential Counterfactuals
DM	Diffusion Model
DPO	Direct Preference Optimization
Eq	Equation
ER	Event Registry
EUS	Expected Utility of Selection
EVOI	Expected Value Of Information
FedAVG	Federated Average
FL	Federated Learning
GFR	several-Group Fixed Random
HEaD	Hallucination Early Detection
HP	Hallucination Prediction
KL	Kullback–Leibler
LLM	Large Language Model
MCTS	Monte-Carlo Tree Search
MIA	Membership Inference Attack
MPL	Multilayer Perceptron
OFM	One-group Fixed subMatrix
OFR	One-group Fixed Random
OSR	One-group reSampled Random
PE	Preference Elicitation
PEAR	Personalized Algorithmic Recourse with Preference Elicitation
PFI	Predicted Final Image
R@K	Recall at Key
RL	Reinforcment Learning
SD	Stable Diffusion
SoTA	State of The Art
SV	Shapley Values
T2I	Text to Image
USR	Unique reSampled Random
VQA	Vision Question Answer

Contents

List of Tables

List of Figures

1 Executive Summary

This deliverable presents the results at M12 for tasks T3.2 (Fairness and Counterfactual Interventions), T3.3 (Cognition-aware hybrid decision-making systems) and T3.4 (Privacy-Preserving Machine Learning). For each task, we describe in detail the methods, results, and outcomes obtained in terms of scientific publications, software, and other materials. What follows is a more detailed list of what can be found in each section of this report.

T3.2 (Fairness and Counterfactual Interventions). The section is devoted to discussing progress made by the partners on topics related to fairness and counterfactual interventions. There are a significant number of topics covered, including bias mitigation in multimodal systems (i.e., image-to-text generation), measuring group fairness using Shapely values, removing toxic contents generated by vision language models, techniques for early detection of hallucination in diffusion models, personalized algorithmic recourse and, eventually, quantifying fairness by using fuzzy logic.

T3.3 (Cognition-aware hybrid decision-making systems). Biases are also covered from a cognition perspective. Indeed, in this section of the report, we discuss the results obtained by partners with a focus on the presence of human-like cognitive biases in AI systems, on the human cognitive biases that emerge when looking at images beautified through the usage of AI-based beauty filters and, eventually, on methods to detect cognitive biases in the news.

T3.4 (Privacy-Preserving Machine Learning). Another aspect covered in this report concerns privacy-preserving machine learning. In this section, we focus on federated learning and we show that membership inference attacks are less effective on clients with larger datasets and more complex models. In addition, a partner investigates how model-agnostic federated learning can help to mitigate such attacks as it allows clients to use models of different complexities to enhance privacy.

2 Introduction

Artificial Intelligence (AI) has rapidly become an integral component of modern society, influencing decision-making processes across diverse domains, from healthcare to finance and beyond. As AI systems become more pervasive, the issues of fairness and bias within these technologies have garnered increasing attention. Ensuring that AI systems operate equitably and transparently is not only an ethical imperative but also critical for maintaining public trust and the integrity of automated decision-making.

The challenge of fairness in AI is multifaceted, encompassing the identification, measurement, and mitigation of biases that may be embedded in these systems. Bias can arise from various sources, including the data used to train models, the design of algorithms, and the societal structures that AI systems reflect and potentially perpetuate. Traditional approaches to addressing bias often focus on well-known categories such as gender, race, and age. However, there is a growing recognition that biases are more numerous and complex than previously considered. This necessitates an open-ended exploration of biases, acknowledging that new and unforeseen biases may emerge as AI technologies evolve.

A critical aspect of achieving fairness in AI is the personalization of interventions. Different individuals and groups may experience the impacts of AI systems differently, requiring tailored approaches to ensure equitable outcomes. This personalization extends beyond simple demographic categorizations, recognizing the unique circumstances and needs of each user. Developing methods for quantifying fairness and bias, such as through advanced metrics and frameworks, is essential for assessing the fairness of AI systems and guiding the implementation of corrective measures.

Moreover, the influence of AI extends into areas such as media and personal aesthetics, where biases can also manifest. The examination of biases in global news reporting highlights the tendency for certain narratives to be emphasized or neglected, which can shape public perception and understanding. Similarly, the use of AI-driven beauty filters often reinforces certain beauty standards, leading to potential biases in how attractiveness is perceived and valued. These areas underscore the pervasive nature of bias and the importance of addressing it across all applications of AI.

In conclusion, this deliverable focuses on the fundamental directions of fairness in AI, emphasizing the need for comprehensive bias identification, personalized interventions, and robust methods for fairness quantification.

3 T3.2 Fairness and Counterfactual Interventions.

3.1 Bias Mitigation in Multimodal Systems

Contributing partners: UNITN

3.1.1 Introduction and Methodology

Text-to-Image (T2I) generation has become increasingly popular, thanks to its intuitive conditioning and the high quality and fidelity of the generated content [9, 10, 11, 12, 5]. Several works extended the base T2I model, unlocking additional use cases, including personalization [13, 14], image editing [15, 16, 17, 18], and various forms of conditioning [19, 20, 21]. This rapid progress urges to investigate other key aspects beyond image quality improvements, such as their fairness and potential bias perpetration [3, 2, 22]. It is widely acknowledged that deep learning models learn the underlying biases present in their training sets [23, 24, 25], and generative models are no exception [22, 3, 2, 26].

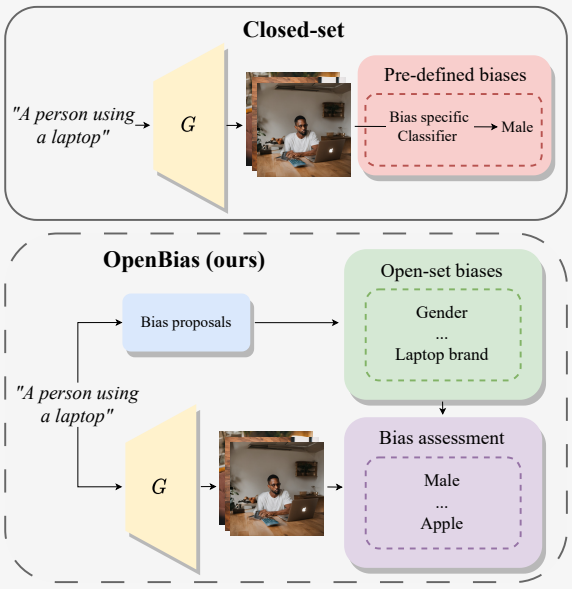


Figure 1. OpenBias discovers biases in T2I models within an open-set scenario. In contrast to previous works [2, 3, 4], our pipeline does not require a predefined list of biases but proposes a set of novel domain-specific biases.

Ethical topics such as fairness and biases have seen many definitions and frameworks [27]; defining them comprehensively poses a challenge, as interpretations vary and are subjective to the individual user. Following previous works [2, 28], a model is considered unbiased regarding a specific concept if, given a context t that is agnostic to class distinctions, the possible classes $c \in \mathcal{C}$ exhibit a uniform distribution. In practice, for a T2I model, this reflects to the tendency of the generator to produce content of a certain class c (e.g., *"man"*), given a textual prompt t that does not specify the intended class (e.g., *"A picture of a doctor"*).

Several works studied bias mitigation in pre-trained models, by introducing training-related methods [29, 30, 31, 32] or using data augmentation techniques [33, 34]. Nevertheless, a notable limitation of these approaches is their dependence on a predefined set of biases, such as gender,

age, and race [22, 2], as well as specific face attributes [3]. While these represent perhaps the most sensitive biases, we argue that there could be biases that remain undiscovered and unstudied.

Considering the example in Fig. 1, the prompt *"A person using a laptop"* does not specify the person's appearance and neither the specific laptop nor the scenario. While closed-set pipelines can detect well-known biases (e.g., gender, race), the T2I model may exhibit biases also for other elements (e.g., laptop brand, office). Thus, an open research question is: *Can we identify arbitrary biases present in T2I models given only prompts and no pre-specified classes?* This is challenging as collecting annotated data for all potential biases is prohibitive.

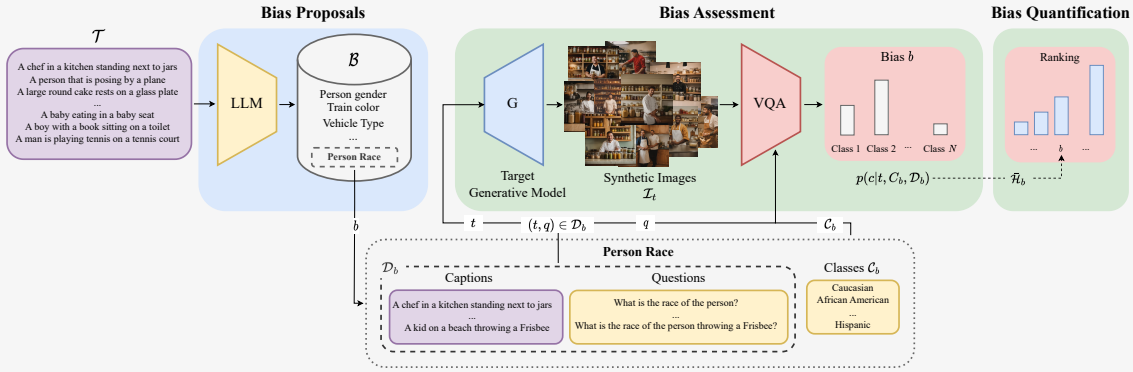


Figure 2. OpenBias pipeline. Starting with a dataset of real textual captions ($\mathcal{T}$) we leverage a Large Language Model (LLM) to build a knowledge base $\mathcal{B}$ of possible biases that may occur during the image generation process. In the second stage, synthesized images are generated using the target generative model conditioned on captions where a potential bias has been identified. Finally, the biases are assessed and quantified by querying a VQA model with caption-specific questions extracted during the bias proposal phase.

"A train zips down the railway in the sun" | *"A photo of a person on a laptop in a coffee shop"* | *"A cop riding a horse through a city neighborhood"*

Figure 3. Novel biases discovered on Stable Diffusion XL [5] by OpenBias.

Toward this goal, we propose *OpenBias* (see Fig.2), the first pipeline that operates in an *open-set scenario*, enabling to identify, recognize, and quantify biases in a specific T2I model without constraints (or data collection) for a specific predefined set. Specifically, we exploit the multi-modal nature of T2I models and create a knowledge base of possible biases given a collection of target textual captions, by querying a Large Language Model (LLM). In this way, we discover specific

biases for the given captions. Next, we need to recognize whether these biases are actually present in the images. For this step, we leverage available Visual Question Answering (VQA) models, directly using them to assess the bias presence. By doing this, we overcome the limitation of using attributes-specific classifiers as done in previous works [2, 3, 35], which is not efficient nor feasible in an open-set scenario. Our pipeline is modular and flexible, allowing for the seamless replacement of each component with newer or domain-specific versions as they become available. Moreover, we treat the generative model as a *black box*, querying it with specific prompts to mimic end-user interactions (i.e., without control over training data and algorithm). We test OpenBias on variants of Stable Diffusion [5, 12] showing human-agreement, model-level comparisons, and the discovery of novel biases.

3.1.2 Experiments

Datasets. We study the bias in two multimodal datasets Flickr 30k [36] and COCO [37]. Flickr30k [36] comprises 30K images with 5 caption per image, depicting images in the wild. Similarly, COCO [37] is a large-scale dataset containing a diverse range of images that capture everyday scenes and objects in complex contexts. We filter this dataset, creating a subset of images whose caption contains a single person. This procedure results in roughly 123K captions. Our choice is motivated by building a large subset of captions specifically tied to people. This focus on the person-domain is crucial as it represents one of the most sensitive scenarios for exploring bias-related settings. Nevertheless, it is worth noting that the biases we discover within this context extend beyond person-related biases to include objects, animals, and actions associated with people.

Figure 4. Novel person-related biases identified on Stable Diffusion XL [5] by OpenBias.

Quantitative Results. Our open-set setting harnesses the zero-shot performance of each component. As in [2], we evaluate OpenBias using FairFace [38], a well-established classifier fairly trained, as the ground truth on gender, age, and race.

Agreement with FairFace. We compare the predictions of multiple SoTA Visual Question Answering models with FairFace. Firstly, we assess the zero-shot performance of the VQA models on synthetic images, performing our comparisons using images generated by SD XL. The evaluation involves assessing accuracy and F1 scores, which are computed against FairFace predictions treated

Model	Gender		Age		Race	
	Acc.	F1	Acc.	F1	Acc.	F1
CLIP-L [39]	91.43	75.46	58.96	45.77	36.02	33.60
OFA-Large [40]	**93.03**	83.07	53.79	41.72	24.61	21.22
mPLUG-Large [41]	**93.03**	82.81	61.37	52.74	21.46	23.26
BLIP-Large [42]	92.23	82.18	48.61	31.29	36.22	35.52
Llava1.5-7B [43, 44]	92.03	82.33	66.54	62.16	55.71	42.80
Llava1.5-13B [43, 44]	92.83	**83.21**	**72.27**	**70.00**	**55.91**	**44.33**

Table 2. VQA evaluation on the generated images using COCO captions. We highlight in gray the chosen default VQA model.

Model	**Flickr 30k** [36]			**COCO** [37]		
	gender	age	race	gender	age	race
Real	0	0.032	0.030	0	0.041	0.028
SD-1.5 [12]	0.072	0.032	0.052	0.075	0.028	0.092
SD-2 [12]	0.036	0.069	0.047	0.060	0.045	0.105
SD-XL [5]	0.006	0.028	0.180	0.002	0.027	0.184

Table 3. KL divergence (↓) computed over the predictions of Llava1.5-13B and FairFace on generated and real images.

as the ground truth. The results are reported in Table 2. Llava1.5-13B emerges as the top-performing model across different tasks, consequently, we employ it as our default VQA model.

Next, we evaluate the agreement between Llava and FairFace [38] on different scenarios. Specifically, we run the two models on real and synthetic images generated with Stable Diffusion 1.5, 2, and XL. We measure the agreement between the two as the KL Divergence between the probability distributions obtained using the predictions of the respective model. We report the results in Table 3. We can observe that the models are highly aligned, obtaining low KL scores, proving the VQA model's robustness in both generative and real settings.

Qualitative Results. We show examples of biases discovered by OpenBias on Stable Diffusion XL. We present the results in a context-aware fashion and visualize images generated from the same caption where our pipeline identifies a bias. We organize the results in three sets and present unexplored biases on objects and animals, novel biases associated with persons, and well-known social biases. We highlight biases discovered on objects and animals in Fig. 3. For example, the model tends to generate *"yellow"* trains or *"quarter horses"* even if not specified in the caption. Furthermore, the model generates laptops featuring a distinct *"Apple"* logo, showing a bias toward the brand.

Next, we display novel biases related to persons discovered by OpenBias. For instance, we unveil unexplored biases such as the *"person attire"*, with the model often generating people in a formal outfit rather than more casual ones. Furthermore, we specifically study *"child gender"* and *"child race"* diverging from the typical examination centered on adults. For example, in Fig. 4 second column, we observe that the generative model links a black child with an economically disadvantaged environment described in the caption as *"a dirt road"*. The association between racial identity and socioeconomic status perpetuates harmful stereotypes and proves the need to consider novel biases within bias mitigation frameworks. Lastly, we show qualitative results on the

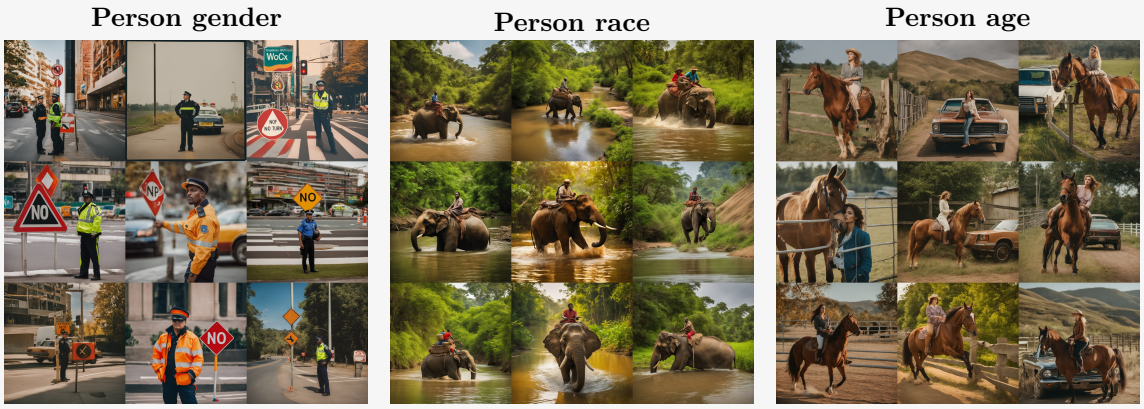


Figure 5. Person-related biases found on Stable Diffusion XL [5] by OpenBias.

well-studied and sensitive biases of *"person gender"*, *"race"*, and *"age"*. In the first column of Fig. 5, Stable Diffusion XL exclusively generates *"male"* officers, despite the presence of a gender-neutral job title. Moreover, we observe a *"race"* bias, with depictions of solely black individuals for *"a man riding an elephant"*. Finally, it explicitly depicts a *"woman"* labeled as *"middle-aged"* when engaged in horseback riding. This context-aware approach ensures a thorough comprehension of emerging biases in both novel and socially significant contexts. These results emphasize the necessity for more inclusive open-set bias detection frameworks.

3.1.3 Conclusion

The main contributions of this work are as follows:

- To the best of our knowledge, we are the first to study the problem of open-set bias detection at large scale without relying on a predefined list of biases. Our method discovers novel biases that have never been studied before.
- We propose OpenBias, a modular pipeline, that, given a list of prompts, leverages a Large Language Model to extract a knowledge base of possible biases, and a Vision Question Answer model to recognize and quantify them.
- We test our pipeline on multiple text-to-image generative models: Stable Diffusion XL, 1.5, 2 [5, 12]. We assess our pipeline showing its agreement with closed-set classifier-based methods and with human judgement.

3.1.4 Relevant Publications

- M. D'Incà, E. Peruzzo, M. Mancini, D. Xu, V. Goel, X. Xu, Z. Wang, H. Shi, and N. Sebe, OpenBias: Open-set Bias Detection in Generative Models, CVPR 2024 [45] Zenodo record: https://zenodo.org/records/11303771

3.1.5 Relevant Software/Datasets/Other Outcomes

The Pytorch implementations can be found in:

- https://github.com/Picsart-AI-Research/OpenBias

3.2 FairShap: Group Fairness via Shapley Values

Contributing partners: ALC

3.2.1 Introduction and Methodology

Machine learning (ML) models are increasingly used to support human decision-making in a broad set of use cases, including in high-stakes domains, such as healthcare, education, finance, policing, or immigration. In these scenarios, algorithmic design, implementation, deployment, evaluation and auditing should be performed cautiously to minimize the potential negative consequences of their use, and to develop fair, transparent, accountable, privacy-preserving, reproducible and reliable systems [46, 47, 48]. To achieve algorithmic fairness, a variety of fairness metrics have been proposed in the literature [49]. Group fairness focuses on ensuring that different demographic groups are treated fairly by an algorithm [50, 51], and individual fairness aims to give a similar treatment to similar individuals [52]. In the past decade, numerous machine learning methods have been proposed to achieve algorithmic fairness [53].

Algorithmic fairness may be addressed in the three stages of the ML pipeline: first, by modifying the input data (*pre-processing*) via e.g. re-sampling, re-weighting or learning fair representations [54, 55]; second, by including a fairness metric in the optimization function of the learning process (*in-processing*) [56, 57]; and third, by adjusting the model's decision threshold (*post-processing*) [50].

From a practical perspective, pre-processing fairness methods tend to be easier to understand for a diverse set of stakeholders, including legislators [58, 59]. Furthermore, to mitigate potential biases in the data, there is increased societal interest in using demographically-representative data to train ML models [60, 61, 62]. However, the vast majority of the available datasets used in real world scenarios are not demographically representative and hence could be biased. Moreover, datasets that are carefully created to be fair lack the required size and variety to train large-scale deep learning models.

In this context, pre-processing fairness methods that focus on modeling and correcting bias on the data emerge as valuable approaches [63]. Methods of special relevance are those that identify the value of each data point not only from the perspective of the algorithm's performance, but also from a fairness perspective [58], and methods that are able to leverage small but fair datasets to improve fairness when learning from large-scale yet biased datasets.

Data valuation approaches are particularly well suited for this purpose. The proposed data valuation methods to date [64] measure the contribution of each data point to the accuracy of the model and use this information as a pre-processing step to improve the performance of the model. However, they have not been used for algorithmic fairness. In this paper, we fill this gap by proposing `FairShap`, an instance-level, data re-weighting method for fair algorithmic decision-making which is model-agnostic and interpretable through data valuation. `FairShap` leverages the concept of SVs [65] to measure the contribution of *each* data point to a pre-defined group *fairness* metric. As the weights are computed on a reference dataset ($\mathcal{T}$), `FairShap` makes it possible to use fair but small datasets to debias large yet biased datasets.

Fig. 6 illustrates the workflow of data re-weighting by means of `FairShap`: first, the weights for each data point x_i in the training set, ϕ_i, are computed based on its contribution to the conditional probabilities of the predicted label given the real label for each group. `FairShap` leverages a reference dataset $\mathcal{T}$ which is either a fair dataset –when available– or the validation set

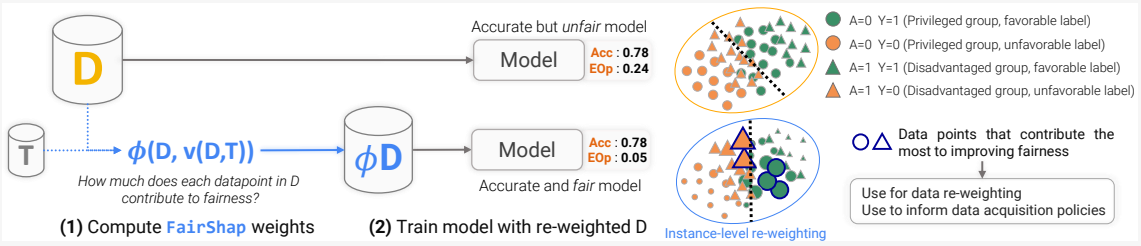


Figure 6. Left: `FairShap`'s workflow. The weights are computed using a reference dataset $\mathcal{T}$, which can be an external dataset or the validation set of D. Right: Illustrative example of `FairShap`'s impact on individual instances and on the decision boundary. Note how `FairShap` re-weighting is able to shift the data distribution yielding a fairer model with similar levels of accuracy.

of the dataset D. Second, once the weights are obtained, the training data is re-weighted. Third, an ML model is trained using the re-weighted data and then applied to the test set.

`FairShap` has several advantages: (1) it is easily interpretable, as it assigns a numeric value (weight) to each data point in the training set; (2) it enables detecting which data points are the most important to improve fairness while preserving accuracy; (3) it makes it possible to leverage small but fair datasets to learn fair models from large-scale yet biased datasets; and (4) it is model agnostic.

Statistical algorithmic fairness depends on the disparity in a model's error rates on different groups of data points in the test set when the groups are defined according to their values of a protected attribute, A. To measure the data valuation for a training data point to the fairness of the model, it is essential to identify the contribution of that training data point to the model's accuracy on the different groups of the test set defined by their protected attribute.

Let $\Phi_{i,j}$ be the contribution of the training point $(x_i, y_i) \in \mathcal{D}$ to the probability of correct classification of the test point $(x_j, y_j) \in \mathcal{T}$. $\Phi_{i,j}$ measures the expected change in the model's correct prediction of j due to the inclusion of i in the dataset, namely:

$$\Phi_{i,j} = \mathbb{E}_{S\sim\mathcal{P}(\mathcal{D}\setminus\{i\})}\left[p(y = y_j|x_j, S\cup\{i\}) - p(y = y_j|x_j, S)\right], \tag{1}$$

and let matrix $\mathbf{\Phi} \in \mathbb{R}^{|\mathcal{D}|\times|\mathcal{T}|}$ be the matrix where each element corresponds to each pairwise train-test point contribution. Leveraging the efficiency axiom, $\phi_i(\text{Acc}) := \mathbb{E}_{j\sim p(\mathcal{T})}[\Phi_{i,j}] = \overline{\mathbf{\Phi}}_{i,:} \in \mathbb{R}$. While a direct implementation of $\Phi_{i,j}$ is very expensive to compute ($O(2^N)$), an efficient implementation ($O(N \log N)$) by [66] is available. It consists of a closed-form solution of $\phi(\text{Acc})$ by means of a deterministic distance-based approach and thus model independent.

`FairShap` considers the family of fairness metrics that are defined by the True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), False Negative Rate (FNR) and their A-Y conditioned versions, namely Equalized Odds (EOdds) and Equal Opportunity (EOp). To obtain fair data valuations, `FairShap` computes $\phi(\mathcal{D}, v)$ based on $\Phi_{i,j}$ and on a reference dataset, $\mathcal{T}$, which can be a small and fair external dataset or a partition of $\mathcal{D}$.

In the following, we derive the expressions to compute the weights of a dataset according to `FairShap` in a binary classification case (i.e., Y is a binary variable) and with binary protected attributes.

Equalized Odds (EOdds) and Equal Opportunity (EOp) are the two group fairness metrics that `FairShap` uses as valuation functions. Given that TPR and TNR are their building blocks, let $\phi_i(\text{TPR})$ and $\phi_i(\text{TNR})$ be two valuation functions that measure the contribution of training point i to the TPR and TNR, respectively. Note that $\text{TPR} = \text{Acc}|_{Y=1}$ and $\text{TNR} = \text{Acc}|_{Y=0}$. Therefore, $\phi_i(\text{TPR})$ corresponds to the expected change in the model's probability of correctly predicting the

positive class when point i is included in the training dataset $\mathcal{D}$, considering all possible training dataset subsets and the distribution of the reference dataset.

$$\begin{aligned}\phi_i(\text{TPR}) :=& \mathbb{E}_{j\sim p(\mathcal{T}|Y=1)}\left[\mathbb{E}_{S\sim\mathcal{P}(\mathcal{D}\setminus\{i\})}\left[p(y=y_j|x_j, S\cup\{i\}) - p(y=y_j|x_j,S)\right]\right] \\ =& \mathbb{E}_{j\sim p(\mathcal{T}|Y=1)}[\Phi_{i,j}] = \overline{\mathbf{\Phi}}_{i,:|Y=1} \in \mathbb{R}\end{aligned} \tag{2}$$

The value for the entire dataset is $\boldsymbol{\phi}(\text{TPR}) = [\phi_0(\text{TPR}), \cdots, \phi_n(\text{TPR})] \in \mathbb{R}^{|\mathcal{D}|}$. $\boldsymbol{\phi}(\text{TNR})$ is obtained similarly but for $Y = y = 0$. In addition, $\phi_i(\text{FNR}) = 1/|\mathcal{D}| - \phi_i(\text{TPR})$ and $\phi_i(\text{FPR}) = 1/|\mathcal{D}| - \phi_i(\text{TNR})$. These four functions fulfill the SV axioms. Intuitively, $\boldsymbol{\phi}(\text{TPR})$ and $\boldsymbol{\phi}(\text{TNR})$ quantify how much the examples in the training set contribute to the correct classification when $y = 1$ and $y = 0$, respectively. Once $\phi_i(\text{TPR})$, $\phi_i(\text{TNR})$, $\phi_i(\text{FPR})$ and $\phi_i(\text{FNR})$ have been obtained, we can compute the `FairShap` weights for a given dataset. However, there are two scenarios to consider, depending on whether the sensitive attribute (A) and the target variable or label (Y) are the same or not.

`FairShap` weights when $A = Y$ In this case, the group fairness metrics (EOp and EOdds) collapse to measure the disparity between TPR and TNR or FPR and FNR for the different values of the actual label [67], Y, which, in a binary classification case, may be expressed as the Equal Opportunity measure computed as EOp := TPR − FPR $\in [-1, 1]$ or its scaled version EOp = (TPR + TNR)/2 $\in [0, 1]$. Thus, the $\phi_i(\text{EOp})$ of data point i may be expressed as

$$\phi_i(\text{EOp}) := \frac{\phi_i(\text{TPR}) + \phi_i(\text{TNR})}{2} \tag{3}$$

`FairShap` weights when $A \neq Y$ This is the most common scenario. In this case, EOp and EOdds use true/false positive/negative rates conditioned not only on Y, but also on A. Therefore, we define $\text{TPR}_{|A=a} = \text{Acc}_{|Y=y,A=a}$, or TPR_a for short, and thus

$$\phi_i(\text{TPR}_a) := \mathbb{E}_{j\sim p(\mathcal{T}|Y=1,A=a)}[\Phi_{i,j}] = \overline{\mathbf{\Phi}}_{i,:|Y=1,A=a} \tag{4}$$

where the value for the entire dataset is $\boldsymbol{\phi}(\text{TPR}_a) = [\phi_0(\text{TPR}_a), \cdots, \phi_n(\text{TPR}_a)]$. Intuitively, $\phi_i(\text{TPR}_a)$ measures the contribution of the training point i to the TPR of the testing points belonging to a given protected group ($A = a$). $\phi_i(\text{TNR}_a)$ is obtained similarly but for $y = 0$. Given EOp := $\text{TPR}_{|A=a} - \text{TPR}_{|A=b}$ and EOdds := $\frac{(\text{FPR}_{A=a} - \text{FPR}_{A=b}) + (\text{TPR}_{A=a} - \text{TPR}_{A=b})}{2}$, then $\phi_i(\text{EOp})$ is given by

$$\phi_i(\text{EOp}) := \phi_i(\text{TPR}_a) - \phi_i(\text{TPR}_b) \tag{5}$$

and $\phi_i(\text{EOdds})$ is expressed as

$$\phi_i(\text{EOdds}) := \frac{(\phi_i(\text{FPR}_a) - \phi_i(\text{FPR}_b)) + (\phi_i(\text{TPR}_a) - \phi_i(\text{TPR}_b))}{2} \tag{6}$$

where their corresponding $\boldsymbol{\phi}(\text{EOp})$ and $\boldsymbol{\phi}(\text{EOdds})$ vectors.

3.2.2 Experiments and Results

Tabular Data. In this section, we consider a common real-life scenario where the target variable Y is not a protected attribute and a single biased dataset $\mathcal{D}$ is used for training, validation, and testing. Thus, the validation set $\mathcal{T}$ is obtained from $\mathcal{D}$ according to the pipeline illustrated in **??**. Given that $A \neq Y$, `FairShap` considers two different valuation functions: $\boldsymbol{\phi}(\text{EOp})$ and $\boldsymbol{\phi}(\text{EOdds})$ as per Eq. (5) and Eq. (6), respectively. Note that in this case the weights assigned to each data point, w_i, are obtained by normalizing $\boldsymbol{\phi}(\text{EOp})$ and $\boldsymbol{\phi}(\text{EOdds})$ following a methodology similar to that described in [68]: $w_i = \phi_i/(\sum_i \phi'_i)|D|$ where $\phi'_i = (\phi_i - \min(\boldsymbol{\phi}))/(\max(\boldsymbol{\phi}) - \min(\boldsymbol{\phi}))$.

Datasets. We test `FairShap` on three commonly used datasets in the algorithmic fairness literature: (1) the German Credit [69] dataset, (2) the Adult Income [70] dataset, and (3) the COMPAS [71] dataset.

Pipeline. The model in all experiments is a Gradient Boosting Classifier (GBC) [72]. Here, the reference dataset $\mathcal{T}$ is the validation set of dataset $\mathcal{D}$. The reported results correspond to the average values of running the experiment 50 times with random splits stratified by sensitive group and label: 70% of the original dataset used for training ($\mathcal{D}$), 15% for the reference set ($\mathcal{T}$) and 15% for the test set. Train, reference and test set are stratified by A and Y such that they have the same percentage of A - Y samples as in $\mathcal{D}$.

Baselines. We compare its `FairShap`'s performance with 6 state-of-the-art algorithmic fairness methods that only *partially* satisfy `FairShap`'s properties: *1. Group RW*: A group-based re-weighting method [54]; *2. Post-processing*: A post-processing approach proposed by [50]; *3. LabelBias:* An in-processing re-weighting technique by [73]; *4. Opt-Pre:* A feature and label transformation-based approach by [74]; *5. IFs:* An Influence Function (IF)-based approach described in [75]; and *6.* $\boldsymbol{\phi}$**(***Acc***)**: A data re-weighting method by means of accuracy-based Shapley Values [64].

Table 4 summarizes the results for the Adult dataset and Fig. 7 summarizes the results on the German Credit and COMPAS datasets.

Adult (Sex|Race)	Acc ↑	M-F1 ↑	EOp ↓	EOdds ↓	Acc ↑	M-F1 ↑	EOp ↓	EOdds ↓
GBC	*.803*±.001	‡.680±.002	‡.451±.004	‡.278±.003	**.803**±.001	‡.682±.002	‡.164±.010	‡.106±.006
Group RW	‡.790±.001	**.684**±.002	*.002*±.009	*.001*±.005	**.803**±.001	‡.683±.002	.010±.009	.010±.005
Postpro	‡.791±.001	†.679±.004	‡.056±.013	‡.034±.007	.802±.001	**.688**±.002	‡.061±.011	‡.042±.006
LabelBias	‡.781±.001	‡.681±.002	‡.065±.011	‡.049±.006	‡.800±.001	*.686*±.002	‡.118±.013	‡.074±.007
OptPrep	‡.789±.001	‡.676±.004	‡.064±.029	‡.037±.017	‡.800±.001	†.685±.002	‡.044±.015	‡.029±.009
IF	‡.787±.002	†.681±.003	‡.159±.037	‡.092±.022	‡.797±.002	†.685±.002	‡.042±.020	‡.031±.012
ϕ(Acc)	**.804**±.001	‡.681±.002	‡.452±.005	‡.279±.003	**.803**±.001	‡.681±.002	‡.161±.011	‡.104±.007
$\boldsymbol{\phi}$**(**EOp**)**	‡.790±.001	**.684**±.002	*.002*±.009	**3e-4**±.005	.802±.001	‡.683±.002	*.009*±.010	*.009*±.005
$\boldsymbol{\phi}$**(**EOdds**)**	‡.790±.001	.683±.002	**8e-4**±.009	*.001*±.005	.802±.001	‡.683±.002	**.007**±.009	**.007**±.005

Table 4. Results on the Adult Income dataset. ***Bold*** *denote the best model and italic the second-best. Statistically significant differences with the best model are denoted by* ‡ *for* $p < 0.01$ *and* † *for* $p < 0.05$.

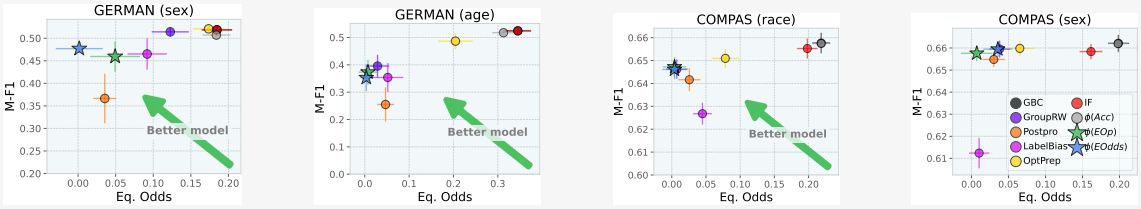
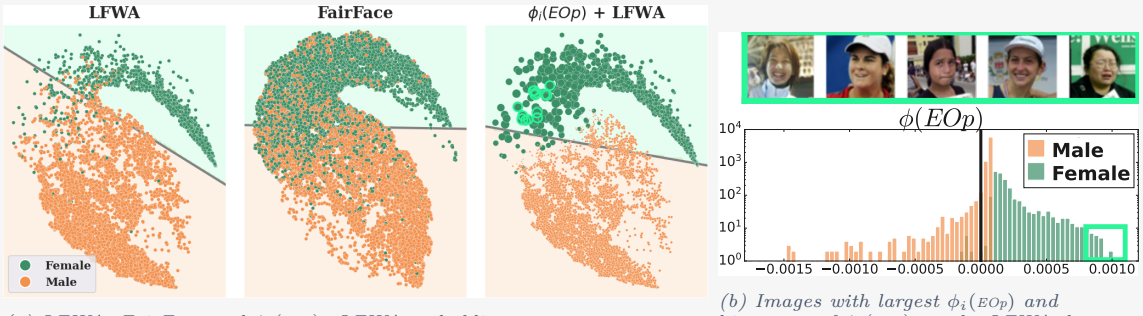


Figure 7. Utility vs. fairness analysis. The models trained with `FairShap` *re-weighting improve fairness while maintaining a competitive level of accuracy compared to the baselines.*

Computer Vision. In this scenario, we focus on a computer vision task to illustrate the versatility of data re-weighting via `FairShap`. In this case, the goal is to predict the sensitive attribute, i.e. $A = Y$. Furthermore, this scenario explores the benefits of leveraging a fair external reference dataset $\mathcal{T}$. The task consists of automatic sex classification from facial images by means of

a deep convolutional network (Inception Resnet V1) using `FairShap` for data re-weighting. Sex (male/female) is therefore both the protected attribute (A) and the target variable (Y).

Datasets. We leverage three publicly available face datasets: CelebA, LFWA [76] and FairFace [77], where LFWA is the training set $\mathcal{D}$ (large-scale and biased) and FairFace is the reference dataset $\mathcal{T}$ (small but fair). The test split in the FairFace dataset is used for testing. CelebA is used to pre-train the Inception Resnet V1 model [78] to obtain the LFWA and FairFace embeddings that are needed to compute the Shapley Values efficiently by means of a k-NN approximation in the embedding space. In the three datasets, sex is a binary variable with two values: male, female.



(a) LFWA, FairFace and $\phi_i(\text{EOp})$+LFWA embeddings.

(b) Images with largest $\phi_i(\text{EOp})$ and histogram of $\phi_i(\text{EOp})$ on the LFWA dataset.

Figure 8. (a) Image embeddings for LFWA (Left), FairFace (Middle) and LFWA with data point sizes $\propto |\phi_i(\text{EOp})|$ (Right). Points with the largest $\phi_i(\text{EOp})$ are highlighted in green. (b) 5 Images with the largest $\phi_i(\text{EOp})$ and histogram of $\phi_i(\text{EOp})$ on the LFWA dataset.

Pipeline. The pipeline to obtain the `FairShap`'s weights in this scenario is depicted in **??** and proceeds as follows: (1) pre-train an Inception Resnet V1 model with the CelebA dataset; (2) use this model to obtain the embeddings of the LFWA and FairFace datasets; (3) compute the weights on the LFWA training set ($\mathcal{D}$) using as reference dataset ($\mathcal{T}$) the FairFace validation partition. (4) fine-tune the pretrained model using the re-weighted data in the LFWA training set according to $\boldsymbol{\phi}$; and (5) test the resulting model on the test partition of the FairFace dataset. The experiment's training details and hyper-parameter setting are described in Appendix D.3.

FairShap Re-weighting. In this case, the group fairness metrics are equivalent and thus, we report results using $\phi_i(\text{EOp})$: $\phi_i(\text{EOp})$ quantifies the contribution of the ith data point (image) in LFWA to the fairness metric (Equal Opportunity) of the model tested on the FairFace dataset.

Baselines. The purpose of this experiment is to illustrate the versatility of `FairShap` in different scenarios rather than to perform an exhaustive comparison with other methods, as previously done with tabular data. Nonetheless, we compare `FairShap` with three baselines: the pre-trained model using CelebA; the fine-tuned model using LFWA without re-weighting; and a data re-weighting approach using $\boldsymbol{\phi}(\text{Acc})$ from [64]. We report two performance metrics: the accuracy of the models in correctly classifying the sex in the images (Acc) and the Equal Opportunity (EOp), measured as $\text{TPR}_M - \text{TPR}_W$, where W is the disadvantaged group (females in this case). We also report the specific TPR for males and females.

Results. The results of this experiment are summarized in Table 5. Note how both re-weighting approaches ($\boldsymbol{\phi}(\text{Acc})$ and `FairShap`) significantly improve the fairness metrics while *increasing the accuracy* of the model. `FairShap` yields the best results **both in fairness and accuracy**. Regarding EOp, the model trained with data re-weighted according to `FairShap` yields improvements of **88%** and **66%** when compared to the model trained without re-weighting (LFWA) and the model trained with weights according to $\boldsymbol{\phi}(\text{Acc})$, respectively. In sum, data re-weighting with `FairShap` is able to leverage complex models trained on biased datasets and improve both their fairness and

accuracy.

Training Set	Acc↑	TPR_W | TPR_M	EOp ↓
FairFace	0.909	0.906 | 0.913	0.007
CelebA	0.759	0.580 | 0.918	0.34
LFWA	0.772	0.635 | 0.896	0.26
$\boldsymbol{\phi}(\text{Acc})$	0.793	0.742 | 0.839	0.09
`FairShap` - $\boldsymbol{\phi}(\text{EOp})$	**0.799**	0.782 | 0.813	**0.03**

Table 5. Performance of the Inception Resnet V1 model tested on the FairFace dataset without and with re-weighting and with binary protected attribute A=Y=sex. The arrows next to the metrics' name indicate if the optimal result of the metric is 0 (↓) or 1 (↑).

To gain a better understanding of the behavior of `FairShap` in this scenario, Fig. 8b (bottom) depicts a histogram of the $\boldsymbol{\phi}(\text{EOp})$ values on the LFWA training dataset. As seen in the Figure, $\boldsymbol{\phi_i}(\text{EOp})$ are mostly positive for the examples labeled as *female* (green) and mostly zero or negative for the examples labeled as *male* (orange). This result makes intuitive sense given that the original model is biased against females, i.e. the probability of misclassification is significantly higher for the images labeled as female than for those labeled as male. Fig. 8b (top) depicts the five images with the largest $\phi_i(\text{EOp})$: they all belong to the female category and depict faces with a variety of poses, different facial expressions and from diverse races.

Note that in this case `FairShap` behaves like a distribution shift method. Fig. 8a shows how $\phi_i(\text{EOp})$ shifts the distribution of $\mathcal{D}$ (LFWA) to be as similar as possible to the distribution of the reference dataset $\mathcal{T}$ (FairFace). Therefore, biased datasets (such as $\mathcal{D}$) may be debiased by re-weighting their data according to $\phi_i(\text{EOp})$, yielding models with competitive performance both in terms of accuracy and fairness. Fig. 8a illustrates how the group fairness metrics impact individual data points: critical data points are those near the decision boundary. This finding is consistent with recent work that has proposed using Shapley Values to identify counterfactual samples [79].

Accuracy fairness trade-off. To further illustrate the impact of `FairShap`'s re-weighting, Fig. 9 depicts the utility-fairness trade-off curves on the three benchmark datasets. We define a parameter α that controls the contribution to the weights of each data point according to `FairShap`, ranging from $\alpha = 0$ (no data re-weighting) to $\alpha = 1$ (weights as given by `FairShap`). Thus, the weights of each data point i are computed as $w'_i = (1-\alpha)\mathbf{1}_{|\mathcal{D}|} + \alpha w_i$ where $\mathbf{1}_n = (1, 1, \ldots, 1) \in \mathbb{R}^n$ is the constant vector and w_i are the weights according to `FairShap`. As shown in Fig. 9, the larger the importance of `FairShap`'s weights, the better the model's fairness. In some scenarios, such as on the German (age) dataset, we observe a utility-fairness Pareto front where the fairest models correspond to $\alpha = 1$ and the best performing models correspond to $\alpha = 0$. Conversely, on the COMPAS (sex) dataset, larger values of α significantly increase the fairness of the model while keeping similar levels of utility (M-F1 and Accuracy).

3.2.3 Conclusion

In this work, we have proposed `FairShap`, an instance-level, model-agnostic data re-weighting approach to achieve group fairness via data valuation using Shapley Values. It is based on the proposed formalization of the pairwise contribution of a training point to the correct classification of a reference point, ϕ_{ij}. We have empirically validated `FairShap` with several state-of-the-art

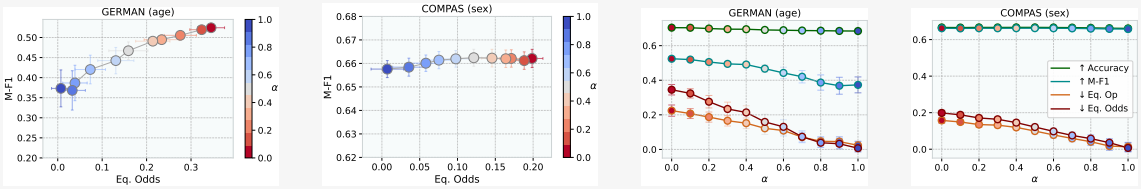


Figure 9. Utility vs fairness trade-off using $\Phi(EOp)$ re-weighting. Left graphs show the MF1-EOdds and the right graphs illustrate the Accuracy, M-F1, EOp and EOdds for increasing values of α.

datasets in different scenarios and using two different types of models (GBCs and deep neural networks). In our experimental results, the models trained with data re-weighted according to `FairShap` delivered competitive accuracy and fairness results. Our experiments also highlight the value of using fair reference datasets ($\mathcal{T}$) for data valuation. We have illustrated the interpretability of `FairShap` by means of histograms and a latent space visualization. We have also studied the accuracy vs. fairness trade-off, the impact of the size of the reference dataset and `FairShap`'s computational cost when compared to baseline models. From our experiments, we conclude that data re-weighting by means of `FairShap` could be a valuable approach to achieve algorithmic fairness. Furthermore, from a practical perspective, `FairShap` satisfies interpretability desiderata proposed by legal stakeholders and upcoming regulations.

3.2.4 Relevant Publications

[80] A. Arnaiz-Rodriguez and N. Oliver. Towards Algorithmic Fairness by means of Instance-level Data Re-weighting based on Shapley Values. ICLR 2024 Workshop on Data-centric Machine Learning Research (DMLR). 2024. https://openreview.net/forum?id=ivf1QaxEGQ

3.2.5 Relevant Software and/or External Resources

- https://github.com/AdrianArnaiz/fair-shap

3.3 Fairness for Toxicity Removal

Contributing partners: UMORE

3.3.1 Introduction and Methodology

Large-scale vision-and-language models, such as CLIP, have demonstrated remarkable efficacy in a variety of tasks, including image classification, cross-modal retrieval, and generation. However, the scale and diversity of the training data required for these models often involve the unsupervised scraping of billions of web items, which can introduce inappropriate content. This issue poses significant risks, particularly when deploying these models in sensitive or trustworthy contexts. Our research addresses this challenge by proposing a novel methodology to enhance the safety of vision-and-language models, specifically targeting the mitigation of NSFW (not safe for work) content.

Our research addresses this issue by proposing a novel methodology aimed at enhancing the safety of vision-and-language models. We specifically focus on reducing the sensitivity of CLIP models to NSFW inputs by disrupting the association between unsafe linguistic or visual items and the corresponding regions of embedding space. This is achieved by fine-tuning a CLIP model

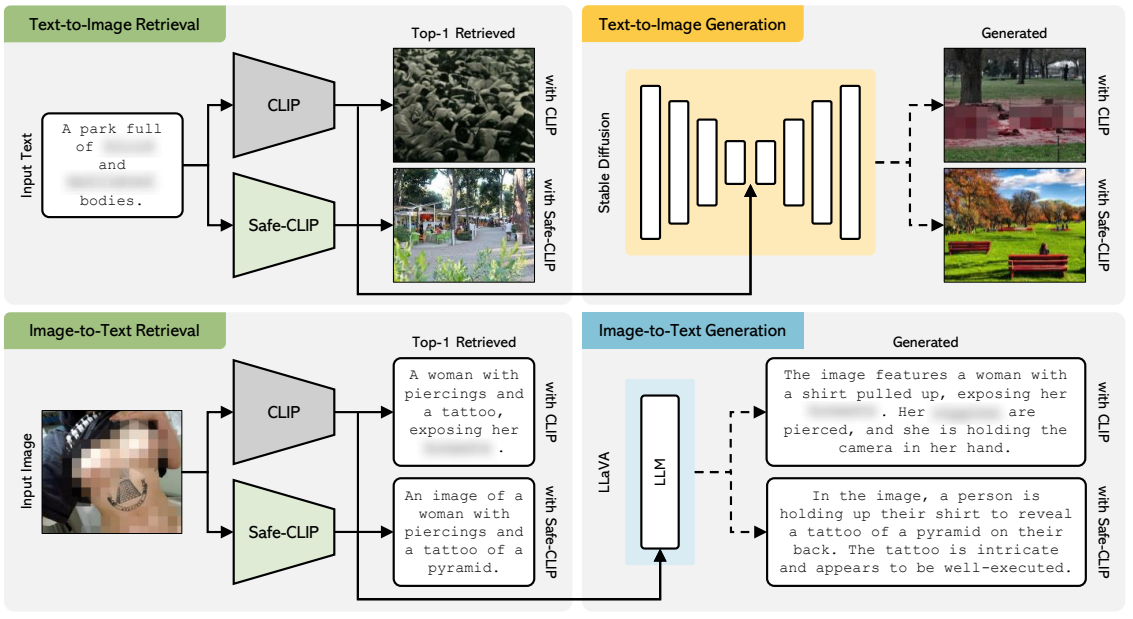


Figure 10. Removing NSFW concepts from CLIP models. Our Safe-CLIP fine-tunes CLIP to make it safer in cross-modal retrieval, image-to-text and text-to-image generation.

on real and synthetic data generated from a large language model trained to convert between safe and unsafe sentences and a text-to-image generator.

The methodology begins with the creation of a toxic language model capable of generating alligned unsafe prompts from safe, visually grounded ones. We fine-tune Llama 2 on manually curated pairs and align it with Direct Preference Optimization (DPO). This fine-tuned model generates unsafe textual content while preserving context alignment and semantic meaning. For generating NSFW images, we use a public open Stable Diffusion XL model available on Hugging Face[1]. Using these models, we generate a dataset of quadruplets consisting of safe and unsafe images and texts to facilitate the fine-tuning of the CLIP embedding space.

Our Safe-CLIP fine-tuning procedure aims to make the embedding space safer by employing a combination of loss functions that redirect inappropriate content while preserving the structure of safe content. Specifically, we define two types of losses: inappropriate content redirection losses and structure preservation losses. To teach the model to ignore inappropriate content, we define the inappropriate content redirection loss using cross-modal cosine similarities between unsafe sentences t_i^* and corresponding safe images v_i, and between unsafe images v_i^* and corresponding safe texts t_i. The loss is formulated as:

$$L_{\text{redir},1} = -\frac{1}{N}\left(\sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{T}(t_i^\star),\mathcal{V}_0(v_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{T}(t_j^\star),\mathcal{V}_0(v_i))/\tau)}+\sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{T}(t_i^\star),\mathcal{V}_0(v_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{T}(t_i^\star),\mathcal{V}_0(v_j))/\tau)}\right. \tag{7}$$
$$\left.+\sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{V}(v_i^\star),\mathcal{T}_0(t_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{V}(v_j^\star),\mathcal{T}_0(t_i))/\tau)}+\sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{V}(v_i^\star),\mathcal{T}_0(t_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{V}(v_i^\star),\mathcal{T}_0(t_j))/\tau)}\right),$$

where τ is a temperature parameter. Additionally, we impose the matching of each unsafe

[1] https://huggingface.co/stablediffusionapi/newrealityxl-global-nsfw

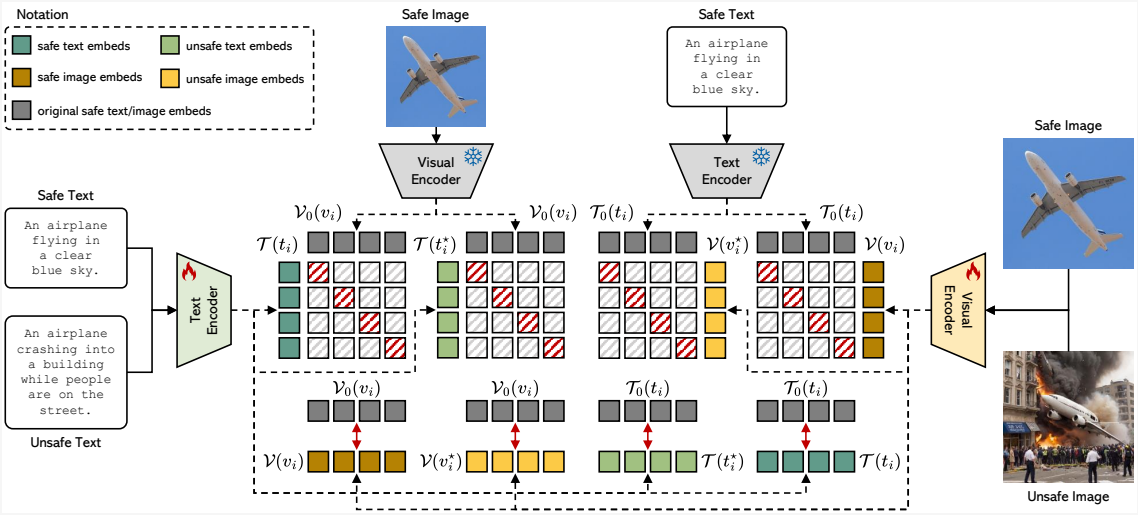


Figure 11. Overview of our Safe-CLIP approach.

embedding with its safe counterpart using the following loss:

$$L_{\text{redir},2} = -\frac{1}{N}\left(\sum_{i=1}^{N}\cos(\mathcal{T}(t_i^\star),\mathcal{T}_0(t_i)) + \sum_{i=1}^{N}\cos(\mathcal{V}(v_i^\star),\mathcal{V}_0(v_i))\right). \tag{8}$$

To preserve the structure of the embedding space for safe content, we introduce a matching loss between the safe embeddings produced by the online networks T and V and those of the original, pre-trained networks T_0 and V_0:

$$L_{\text{pres},1} = -\frac{1}{N}\left(\sum_{i=1}^{N}\cos(\mathcal{T}(t_i),\mathcal{T}_0(t_i)) + \sum_{i=1}^{N}\cos(\mathcal{V}(v_i),\mathcal{V}_0(v_i))\right). \tag{9}$$

We also include a contrastive loss to maintain the original cross-modal relationships between safe visual and textual embeddings:

$$\begin{aligned} L_{\text{pres},2} = -\frac{1}{N}\Bigg(&\sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{V}_0(v_i),\mathcal{T}(t_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{V}_0(v_i),\mathcal{T}(t_j))/\tau)} + \sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{V}_0(v_i),\mathcal{T}(t_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{V}_0(v_j),\mathcal{T}(t_i))/\tau)} \\ &+\sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{T}_0(t_i),\mathcal{V}(v_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{T}_0(t_i),\mathcal{V}(v_j))/\tau)} + \sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{T}_0(t_i),\mathcal{V}(v_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{T}_0(t_j),\mathcal{V}(v_i))/\tau)}\Bigg). \end{aligned} \tag{10}$$

The overall loss function used to fine-tune the network is a weighted sum of these four loss functions.

3.3.2 Experiments and Results

This section presents a comprehensive evaluation of Safe-CLIP, focusing on cross-modal retrieval, robustness testing with real NSFW images, and generative performance in text-to-image and image-to-text generation. We employed a mix of synthetic and real datasets to assess the model's performance and safety enhancements.

Table 6. Retrieval results on the ViSU test set. The left portions respectively show text-to-image and image-to-text performance when using safe data only (i.e. $\mathbf{V}$ *and* $\mathbf{T}$*). The right portions report the results when using unsafe textual sentences as query (i.e.* $\mathbf{T}^\star$*) and the merging of safe (i.e.* $\mathbf{V}$*) and unsafe images (i.e.* $\mathbf{V}^\star$*) as retrievable items, or when using unsafe visual queries (i.e.* $\mathbf{V}^\star$*) and the merging of safe (i.e.* $\mathbf{T}$*) and unsafe sentences (i.e.* $\mathbf{T}^\star$*) as retrievable items.*

	Text-to-Image ($\mathbf{T}$-to-$\mathbf{V}$)			**Image-to-Text** ($\mathbf{V}$-to-$\mathbf{T}$)			**Text-to-Image** ($\mathbf{T}^\star$-to-$\mathbf{V} \cup \mathbf{V}^\star$)			**Image-to-Text** ($\mathbf{V}^\star$-to-$\mathbf{T} \cup \mathbf{T}^\star$)		
Model	R@1	R@10	R@20	R@1	R@10	R@20	R@1	R@10	R@20	R@1	R@10	R@20
CLIP (ViT-L) [39]	36.8	71.6	81.5	39.8	74.2	83.5	2.0	24.8	33.2	4.5	32.9	40.6
DataComp-1B (ViT-L) [83]	46.7	79.7	87.4	47.0	81.3	88.9	1.6	28.1	35.6	5.5	37.5	44.9
w/o inap. content redirection	49.9	83.7	90.3	48.1	83.6	90.5	1.6	30.4	40.1	6.1	35.2	42.6
w/o negative cosine similarities	41.9	78.5	87.3	41.5	77.8	86.9	**8.2**	46.0	56.6	13.7	60.4	68.2
Safe-CLIP	45.9	81.8	89.7	45.3	82.3	89.7	8.0	**46.9**	**58.0**	**19.1**	**62.9**	**71.1**

Dataset. The primary dataset used for training and evaluation is ViSU, containing 165,000 quadruplets of safe and unsafe images and texts. The unsafe data was generated using a conditioned NSFW image generator based on the public open SDXL model available on Hugging Face[2]. For robustness testing and evaluation, we also utilized real NSFW images sourced from NudeNet[81], images crawled from the web using NSFW data source URLs[3], and the Socio-Moral Image Database (SMID)[82]. The I2P dataset[6], comprising 4,703 textual prompts with various categories of inappropriate content, was used to evaluate text-to-image generation.

Evaluating the Safe-CLIP Embedding Space.
Results on ViSU test set. To evaluate the retrieval performance of Safe-CLIP, we firstly considered image-to-text and text-to-image retrieval in a safe-only setting, where we do not have any inappropriate content in both visual and textual data. This is important to assess whether the properties of the original CLIP embedding space are preserved when employing our fine-tuning strategy. In this case, query elements are represented by the safe images of the test set (which, with a slight abuse of notation, we refer to as $\mathbf{V}$) for the image-to-text setting and the safe textual items (referred to as $\mathbf{T}$) for the text-to-image one. Moreover, we considered text-to-image and image-to-text retrieval when using unsafe texts as queries (referred to as $\mathbf{T}^\star$) and both safe and unsafe images as retrievable items and when using unsafe images as queries (*i.e.* $\mathbf{V}^\star$) and both safe and unsafe texts as retrievable items.

Retrieval results on the ViSU test set are reported in Table 6, comparing the proposed Safe-CLIP model with the original CLIP architecture, a CLIP model trained on the DataComp dataset [83], which has undergone NSFW content cleaning, and two different baselines. Specifically, we considered a variant of our approach in which we removed the two negative cosine similarity losses (*i.e.* Eq. 8 and 9), and a model trained with safe data only (*i.e.* removing the loss functions for inappropriate content redirection reported in Eq. 7 and 8). Results are reported in terms of Recall@k (R@k) with $k = 1, 10, 20$, that measures the percentage of times the visual or textual item associated to the query is retrieved among the top-k elements. When using unsafe sentences as queries, for each element we considered the *safe* image associated with the given text as the corresponding visual element. Symmetrically, when using unsafe images as queries, for each element we considered the *safe* text associated with the given image as the ground-truth item. Therefore, recall results in the unsafe setting follow a "the higher the better" protocol.

[2] stablediffusionapi/newrealityxl-global-nsfw
[3] https://github.com/EBazarov/nsfw_data_source_urls

Figure 12. Top-1 images (left) and text (right) retrieved using the original CLIP model and our Safe-CLIP, when NSFW texts and images are employed as query.

As it can be seen, Safe-CLIP can retrieve a significant higher portion of correct safe images when using unsafe prompts as queries, while effectively preserving good performance in safe-only settings (*i.e.* **V**-to-**T** and **T**-to-**V**). Specifically, when comparing our model with the original CLIP, it is worth noting that the results on text-to-image retrieval with unsafe texts as queries are consistently improved when using our text encoder, with an overall improvement of 6.0 points in terms of R@1, and the same applies for image-to-text retrieval which showcases an improvement of 14.6 R@1 points. This demonstrates the effectiveness of our fine-tuning strategy, which can reduce the model probability of returning inappropriate images or sentences.

Robustness on real NSFW images. To further analyze the safety degree of the Safe-CLIP embedding space, we performed text-to-image and image-to-text retrieval using real NSFW images as visual items. Specifically, we selected inappropriate visual content from three different sources: *(i)* a portion of data used to train the NudeNet classifier, *(ii)* images crawled from the web using NSFW data source URLs[4], and *(iii)* images from the Socio-Moral Image Database (SMID) [82]. While the first two sources exclusively contain nudity and pornography images, the third one includes more varied types of inappropriate images representing negative concepts such as, for example, *harm*, *inequality*, *discrimination*, and *unfairness*. Overall, we randomly sampled 1,000 images from each of the NSFW data sources, selecting only those representing unsafe concepts for the SMID dataset. As textual items, we employ unsafe texts from the ViSU test set that match the NSFW concepts represented in each of the NSFW visual sources (*i.e. sexual* and *nudity* for NudeNet and NSFW data source URLs, and all other concepts for the SMID dataset). For both I2T and T2I, we employ a set of 10k randomly selected visual or textual distractors, randomly selected from the LAION-400M dataset [84].

Results show that Safe-CLIP consistently reduces the percentage of retrieved NSFW items for all three NSFW dataset sources. In particular, the percentage of retrieved NSFW visual content is reduced from over 55% to below 10% for NudeNet and NSFW URLs, and from 47.8% to 16.7% for SMID. Similarly, the retrieval of NSFW textual content dropped from over 60% to approximately 25% for NudeNet and NSFW URLs, and from 41.4% to 34.5% for SMID. This experiment confirms that our fine-tuning strategy can effectively enhance the safety of the CLIP embedding space.

Qualitative results. Fig. 12 reports qualitative retrieval results in the same aforementioned setting. Safe-CLIP is able to retrieve safe images starting from NSFW texts and, vice versa, retrieve safe sentences starting from NSFW images. Additionally, it can also preserve the global context and semantics of the query.

Safe-CLIP for Text-to-Image Generation.

Results on I2P and ViSU test set. We then validated the effectiveness of the Safe-CLIP text encoder when applied in a text-to-image generative model. Specifically, we emploied Stable

[4] https://github.com/EBazarov/nsfw_data_source_urls

Figure 13. Images generated from unsafe prompts with Stable Diffusion, employing the original CLIP model, negative prompts, SLD-Strong [6], and our Safe-CLIP.

Diffusion v1.4 [85], replacing the standard CLIP text encoder used in Stable Diffusion with our fine-tuned version. Moreover, we also applied Safe-CLIP in combination with other NSFW removal strategies. In particular, we considered a version of Stable Diffusion with negative prompts and the recently proposed Safe Latent Diffusion (SLD) approach [6], which employs different levels of safety guidance (SLD-Weak, SLD-Medium, and SLD-Strong) to limit the generation of inappropriate images. For this experiment, we generated five images for each textual prompt using different random seeds and computed the probability of generating inappropriate images detected by two NSFW classifiers. Following [6], we employ Q16 [86] and NudeNet [81].

Results indicate that Safe-CLIP significantly reduced the probabilities of generating NSFW images when using textual prompts from both I2P and ViSU datasets. Specifically, the application of Safe-CLIP to the standard Stable Diffusion model decreased the probability of generating inappropriate content by 13.5 percentage points for I2P prompts and by 22.6 percentage points for NSFW texts from ViSU. Similar improvements were observed when combining Safe-CLIP with other NSFW removal strategies. For instance, applying Safe-CLIP in conjunction with negative prompts or SLD-Weak reduces the generation of NSFW images to below 5% across various categories.

Qualitative results. Samples of generated images are shown in Fig. 13, comparing results generated by Safe-CLIP applied to Stable Diffusion with images generated by SLD-Strong [86], Stable Diffusion with negative prompts, and the Stable Diffusion original version. Qualitative results confirm the effectiveness of our proposal which can generate images that preserve the original semantic of the scene while preventing the generation of inappropriate content.

Safe-CLIP for Image-to-Text Generation.
Finally, we assessed the capabilities of the Safe-CLIP visual encoder when applied to an existing multimodal LLM [87]. We employed LLaVA [88] based on LLama 2-13B-Chat, prompted by asking it to describe a given image. We used real NSFW images from three different sources: NudeNet, NSFW URLs, and the Socio-Moral Image Database (SMID). The evaluation metrics included the percentage of NSFW generated texts measured with GPT-3.5 and the toxicity degree computed using the Perspective API.

Results indicate that Safe-CLIP significantly reduced the probability of generating inappropriate textual sentences. Specifically, when using NudeNet images, the percentage of NSFW text generated drops from 62.6% to 26.7%, and the toxicity score decreased from 38.6 to 16.5. For images from NSFW URLs, the percentage of NSFW text generated reduced from 46.8% to 19.4%, and the toxicity score fell from 24.9 to 10.8. Similarly, for SMID images, the NSFW text generation probability dropped from 22.2% to 11.7%, with the toxicity score decreasing from 4.7 to 3.7. These results demonstrate the effectiveness of Safe-CLIP in enhancing the safety of image-to-text

generation models.

3.3.3 Conclusion

Safe-CLIP is a fine-tuning technique designed to make models like CLIP safer and less sensitive to NSFW content. To achieve this, we utilized a large generated synthetic dataset containing both safe and unsafe images along with corresponding safe and unsafe captions. This dataset was used to fine-tune CLIP with specific loss functions aimed at redirecting unsafe content while preserving the overall structure of the CLIP embedding space.

To improve the quality of dataset creation, we employed models presented in the literature and fine-tuned a LLM to generate alligned pairs of safe and unsafe captions.

Experimental results demonstrate the effectiveness of our approach across a variety of tasks. We conducted experiments in cross-modal retrieval, image-to-text generation, and text-to-image generation, all of which highlighted the suitability of Safe-CLIP.

3.3.4 Relevant Publications

[89] S. Poppi et al. Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models. European Conference on Computer Vision (ECCV), 2024. https://arxiv.org/abs/2311.16254

3.3.5 Relevant Software and/or External Resources

- https://github.com/aimagelab/safe-clip
- https://huggingface.co/datasets/aimagelab/ViSU-Text
- https://huggingface.co/aimagelab/safeclip_vit-l_14

3.4 Hallucination Early Detection in Diffusion Models

Contributing partners: UMORE

3.4.1 Introduction and Methodology

Generative models, while progressing, often hallucinate when generating multiple or "long-tail" objects [90, 91, 92]. Furthermore, they frequently have shortcomings in ensuring that attributes, counts, and semantic object relations are correctly generated. This is especially problematic when they are tasked with rendering scenes involving multiple objects [93, 94], where diffusion patterns often produce inconsistencies, significantly impacting the quality of the output [95].

In our study, we assess the performance of Stable Diffusion v1.4 (SD1.4) and Stable Diffusion v2 (SD2) [85] when tasked with generating images from prompts that request the presence of a different number of objects. Our analysis reveals that the probability of a perfect generation – capturing all requested objects – with four objects is only 26.96% for SD1.4 and improves to 30.61% for SD2, as illustrated in Fig. 14(a). The effectiveness of an image generation diffusion model depends strongly on the choice of the initial seed, which establishes the starting latent noise and guides the model through the latent space [96, 97, 93, 98, 92]. This aspect is underlined by our results: in scenarios with four objects, at least one among the 11 seeds we tested leads to a perfect generation in 79.87% of the prompts. This significant percentage highlights the central role of seed selection in overcoming the unpredictability and variability associated with these models, indicating substantial opportunities for improving generative accuracy. While some attempts have been made to develop automatic evaluation metrics for image generation [99, 100], these still fail

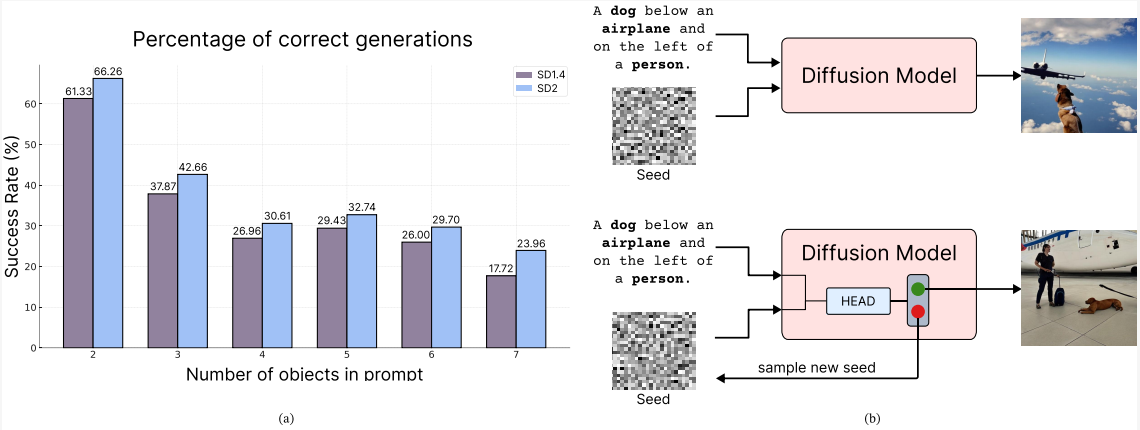


Figure 14. a) Proportion of perfect generations, i.e. images featuring all requested objects, as a function of the number of objects requested in the prompt (InsideGen dataset).
b) Overview of the HEaD pipeline: during the generation process, HEaD assesses whether all designated objects will be accurately represented in the final image, determining if the generation process should continue or be restarted with a different seed.

at ensuring a sufficiently fast and reliable evaluation. Further, even in the presence of a reliable assessment of the generated image, the generation process should be repeated multiple times until reaching a correct generation. In this report, we instead take a different path and explore solutions for an early (*i.e.*, at early stages of the diffusion process) detection of the risk of hallucination. We use this in combination with a procedure to abort or restart generation with another seed to save time and improve the final quality.

To this aim, we focus on specific hallucinations: the omission in the generated image of one or more target objects requested in the textual prompt.

Our approach, termed **HEaD** – Hallucination Early Detection, is the first approach designed to enhance both efficiency and accuracy of generative DMs. HEaD incorporates the use of intermediate cross-attention maps to examine the relationship between the prompt and the internal attention layers of the model, along with the *Predicted Final Image* (PFI) – a prediction of the expected outcome at intermediate stages of the generation process. The combination of PFIs and cross-attention maps allows for the early identification of potential errors by predicting the inclusion or exclusion of objects requested by the initial prompt. By preemptively detecting these anomalies, HEaD hints at stopping the generation diffusion process, thereby conserving resources and reducing the time spent on generating images that would not ultimately meet quality standards.

HEaD is developed through the training of an Hallucination Prediction (HP) network on a dataset featuring both accurate and hallucinated images. Specifically, we generated **InsideGen**, a dataset that encompasses 45,000 images generated with SD2 and SD1.4 [85], saving cross-attention maps and PFIs at intermediate steps. Leveraging on this dataset, HEaD detector is designed to integrate seamlessly with all existing Diffusion Models, enhancing their ability to reliably produce images that encompass all requested objects. This improvement is evident when HEaD is applied to SD1.4 and TokenCompose [1], even with datasets and prompt categories that are entirely different from those in InsideGen.

HEaD Architecture. The primary goal of HEaD is to detect and preemptively interrupt faulty generative processes. Our approach is designed to verify the consistency between the input textual guidance and the expected output *during* the diffusion image generation process to save computa-

tional time and ensure correct generation. A distinctive feature of our approach is its capability to conduct this verification at a specific intermediate timestep of the diffusion pipeline. To this aim, we designed an Hallucination Prediction (HP) network to detect the risk of hallucination. If an hallucination risk is detected, our approach forces the generation process to restart using a different initial seed that might lead to better results. In the following, we illustrate the proposed HEaD approach at inference time to streamline the generation process and, as a result, enable automatic quality assessment of the final output.

Cross-Attention Maps and PFI Extraction. Let us consider a prompt y containing a set of target objects O to be generated in the image. The process of extracting these target objects from the prompt can be formalized as:

$$O = \text{TOE}(y), \tag{11}$$

where $\text{TOE}(\cdot)$ represents a Target Object Extraction function. Here, the term "objects" refers to words in the prompt directly associated with discernible elements in the image, for which we will extract the corresponding cross-attention maps. While our current methodology primarily focuses on objects, it holds the capability for future expansion to include a wider spectrum of visual concepts (such as attributes or colors), going beyond the limits of object-based extraction.

We then define the *critical timestep*, denoted as $\mathcal{T}$, as a specific step in the diffusion process where cross-attention maps for each object ($A_{O,\mathcal{T}}$) and the Predicted Final Image ($\text{PFI}_{\mathcal{T}}$) are extracted. These elements are then used as inputs for our Hallucination Prediction network.

In particular, for each object $o \in O$, the cross-attention map $A_{o,\mathcal{T}}$ is obtained by applying the function $a(\cdot)$, as defined in Eq. **??**. $\text{PFI}_{\mathcal{T}}$, instead, represents the prediction of the expected outcome at the end of the generation process, using only information available at timestep $\mathcal{T}$. In particular, the scheduler projects the latents at $\mathcal{T}$ to the final step, and the decoder translates these predicted latents into the image space. Formally, this process can be defined as follows:

$$\begin{aligned} \epsilon_{\mathcal{T}} &= \epsilon_\theta(z_{\mathcal{T}}, \mathcal{T}) \\ z_0^{\mathcal{T}} &= \Delta(z_{\mathcal{T}}, \epsilon_{\mathcal{T}}, \mathcal{T}, 0) \\ \text{PFI}_{\mathcal{T}} &= D(z_0^{\mathcal{T}}) \end{aligned} \tag{12}$$

where $\epsilon_{\mathcal{T}}$ represents the predictive noise obtained from the UNet model at critical timestep $\mathcal{T}$. The function Δ updates the latents $z_{\mathcal{T}}$ to the predicted latents at the final timestep, denoted as $z_0^{\mathcal{T}}$. Finally, the decoder D translates these predicted final latents into the Predicted Final Image, $\text{PFI}_{\mathcal{T}}$.

As an additional input for the model, we extract an embedding vector for each object. This vector is produced using CLIP [39] from the text of the requested objects extracted from the prompt y. Specifically, for each object, we have $v_o = \text{CLIP}_{\text{Text}}(y_o)$ obtained by applying the CLIP model to the textual representation y_o of each object.

$\text{PFI}_{\mathcal{T}}$, the attention maps $A_{O,\mathcal{T}}$ and the textual embeddings enable the HP network to meticulously assess and predict the presence of specified objects in the final generated image, ensuring a coherent and accurate output aligned with the initial textual guidance.

3.4.1.1 Hallucination Prediction Network.

During the evaluation phase, the Hallucination Prediction network process a single object o at a time, taking as input its cross-attention map $A_{O,\mathcal{T}}$, the textual embedding vector of the object v_o and the predicted final image, $\text{PFI}_{\mathcal{T}}$. The network then outputs a binary prediction indicating the presence or absence of that target object in the final image. Formally, the process can be written as

$$H_o = \text{HP}(A_o, \text{PFI}_{\mathcal{T}}, v_o), \tag{13}$$

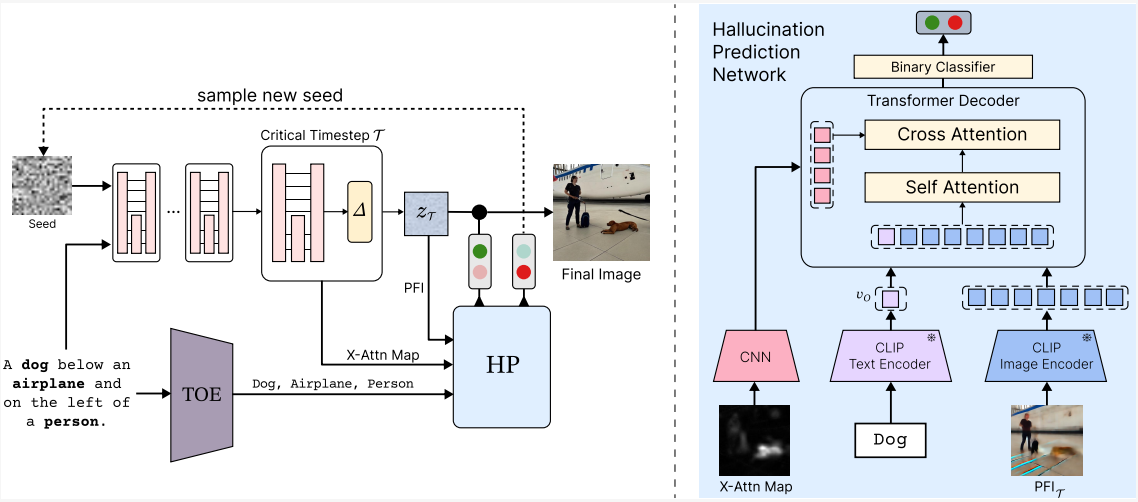


Figure 15. Overview of the hallucination detection process and detail of the HEaD network. The process starts with extracting subjects from the prompt using the TOE model. At timestep $\mathcal{T}$ of the diffusion process, cross-attention maps and PFI are produced. HEaD network processes all inputs using specific feature extractors, and combines them using a Transformer-based decoder to predict whether generation should continue or not.

where H_o is the binary prediction for object o. An image is considered complete if $\forall o \in O, H_o = 1$, otherwise, at inference time, the process must restart.

The HP architecture consists of a Transformer model with cross-attention layers that integrates the two input streams (PFI$_{\mathcal{T}}$, $A_{o,\mathcal{T}}$) through self-attention and cross-attention mechanisms. Initially, all streams are elaborated separately. The PFI is subjected to processing via the visual CLIP backbone with the extraction of features from the last attention layer. This is followed by the concatenation of the textual token v_o to this sequence. Simultaneously, a three-layer convolutional network, processes the cross-attention map.

Following these operations, both outputs are adjusted to the embedding dimension of the transformer model, predetermined at 192, by employing 1D convolution. This stage precedes the implementation of a self-attention mechanism on the stream originating from CLIP, followed by a cross-attention layer that combines the CLIP features (acting as the query) with the ones obtained from the Cross-Attention maps (serving as keys and values). Skip connections are added following the classical architecture of the Transformer decoder [101]. This attention block is consistently applied 12 times, concluding with a classification head that takes as input the activation corresponding to the first element of the input sequence, *i.e.* v_o. The model architecture is shown in Fig. 15

Throughout the training phase, the visual and textual backbones of CLIP are kept frozen. This approach guarantees substantial generalization capabilities towards objects not encompassed within the training dataset, thereby accommodating applications in diverse and unpredictable environments.

The reliability of the HP network is critical to prevent unnecessary terminations of the generation and ensures that objects that would have been present are not prematurely discarded.

3.4.2 Experiments and Results

To compare the generation quality, we followed the evaluation protocol introduced in [1]. Specifically, in this benchmark, the text-to-image model is prompted to generate 5 subjects (*i.e.* A,B,C,D,E)

Method	COCO			
	MG2	MG3	MG4	MG5
SD 1.4[85]	$90.72_{1.33}$	$50.74_{0.89}$	$11.68_{0.45}$	$0.88_{0.21}$
SD 1.4[85] + HEaD	$\mathbf{95.44}_{0.67}$	$\mathbf{65.03}_{1.76}$	$\mathbf{19.53}_{1.73}$	$\mathbf{1.78}_{0.24}$
Composable Diffusion [102]	$63.33_{0.59}$	$21.87_{1.01}$	$3.25_{0.45}$	$0.23_{0.18}$
Layout Guidance [103]	$93.22_{0.69}$	$60.15_{1.58}$	$19.49_{0.88}$	$2.27_{0.44}$
Structured Diffusion [94]	$90.40_{1.06}$	$48.64_{1.32}$	$10.71_{0.92}$	$0.68_{0.25}$
Attend-and-Excite [92]	$93.64_{0.76}$	$65.10_{1.24}$	$28.01_{0.90}$	$\mathbf{6.01}_{0.61}$
TokenCompose [1]	$\mathbf{98.08}_{0.40}$	$76.16_{1.04}$	$28.81_{0.95}$	$3.28_{0.48}$
TokenCompose [1] + HEaD	$97.61_{0.40}$	$\mathbf{81.27}_{1.40}$	$\mathbf{35.33}_{1.97}$	$4.93_{0.57}$

Table 7. SOTA comparison on the evaluation protocol introduced in [1]. Results without HEaD are taken from Wang et al.[1].

with the caption *A photo of A, B, C, D, and E*. Subsequently, an open-vocabulary object detector [104] is asked to detect the presence of the subjects in the generated images. In Table 7, we report the performance of HEaD and competitors following the COCO subset prompts. This split includes 80 objects extracted from COCO [105] combined as previously defined, building 1000 different prompts. Considering the capability of each model to potentially introduce hallucinations, the generated images may feature between one to five subjects. The metric MG-N is utilized to quantitatively ascertain the count of generated images that accurately portray a minimum of N subjects as requested in the initial prompt. Results are reported with mean and standard deviation over 10 seeds following the original implementation.

To promote a fair comparison, we impose a limitation on the maximum number of restarts for HEaD, specifically setting this limit to 5 iterations. This constraint is designed to standardize the inference time, facilitating a direct comparison with alternative methodologies. In instances where a correct image (*i.e.*, an image encapsulating all designated objects) fails to materialize within these iterations, the seed with the highest number of objects predicted by HEaD is chosen for the generation. Additionally, the critical timestep $\mathcal{T}$ is set to 25. The defined HEaD methodology proves beneficial in enhancing the overall quality of the output.

As illustrated in Table 7, the implementation of HEaD enhances the detection of object presence in comparison to SD1.4 [85]. Specifically, an average increase of 14.29% in the detection of three objects is observed when HEaD is integrated with SD1.4 compared to the raw SD. Additionally, improvements of 7.85% and 0.9% are noted for MG4 and MG5, respectively. When compared to Composable Diffusion [102], Layout Guidance [103], and Structured Diffusion [94]. HEaD with SD1.4 surpasses them in nearly all the MG categories. Specifically, for MG3 metric, SD1.4 equipped with HEaD obtains a gain of 43.16%, 4.8%, and 16.39% for respectively Composable, Layout Guidance, and Structured Diffusion.

Significantly, HEaD was initially trained using the model SD2, which demonstrates the capability of the model to adapt to various diffusion models beyond the one it was originally trained on. This adaptability feature is further corroborated by the performance enhancements achieved when HEaD is utilized in conjunction with TokenCompose [1]. In this scenario HEaD obtains an increase of 5.11%, 6.52%, and 1.65% in MG3, MG4, and MG5 when compared to TokenCompose, providing state-of-the-art performance in MG3 and MG4. While Attend-and-Excite [92] records the best results for MG5, TokenCompose equipped with HEaD obtains comparable results.

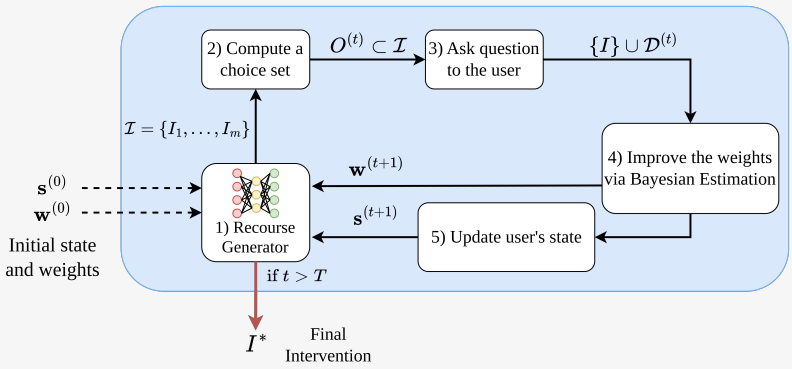


Figure 16. Overview of PEAR. (1) Given the initial state $\boldsymbol{s}^{(0)}$ of the user and weights $\boldsymbol{w}^{(0)}$, PEAR computes a pool of candidate interventions achieving recourse. (2) A choice set $O^{(t)}$ is selected from the pool and presented to the user. (3) The user picks their preferred intervention from the set. (4) An improved estimate of the weights $\boldsymbol{w}^{(t+1)}$ is computed using this feedback, and (5) the user's state $\boldsymbol{s}^{(t+1)}$ is updated. After T rounds, the estimated weights are used to compute a final intervention I^*.

3.4.3 Conclusion

In this work, we have proposed HEaD, a state-of-the-art methodology designed to significantly improve the efficiency and accuracy of image generation processes using Diffusion Models. The key to our innovation lies in the integrated use of cross-attention maps, Predicted Final Image, and textual data to predict the outcome of the image generation process. HEaD has been shown to improve the generation fidelity of the requested objects inside the final image. One of the key messages we wish to convey through our work is that HEaD can be effectively applied "as is" to other diffusion models to ensure the presence of targeted objects in the final image. The internal operations of the diffusion model are not affected by the functionality of HEaD. Furthermore, we created InsideGen, a dataset that we plan to make available to the research community. This resource will facilitate further investigation into how the internal generation data can be leveraged to refine the image generation process as a whole. The HEaD approach could be further expanded not only evaluating the presence/absence of objects but also the relationships of the objects or the aesthetic aspects of the objects themselves.

3.5 Personalized Algorithmic Recourse

Contributing partners: FBK, UNITN

3.5.1 Introduction and Methodology

Automated decision support systems are increasingly employed in high-risk decision tasks to empower human decision-makers and improve the quality of their decisions. Example applications include bail requests [106], loan approvals [107], job applications [108], and prescription of medications and treatments [109]. Despite their promise, often these systems are opaque – meaning that users, and even engineers, have trouble understanding and controlling their decision process – and provide no means for overturning unwanted outcomes, such as denied loan requests. One way of addressing these issues is through the lens of *Algorithmic Recourse* (AR) [110, 111]. In AR, given an undesirable machine-generated decision, the goal is to identify a sequence of actions – or *interventions* for short – that once implemented by the user overturns said decision, for instance,

changing job or obtaining a master's degree. Motivated by this, a number of approaches have been recently proposed for computing AR [112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123].

It is critical that the suggested recourse plans are not too difficult or expensive to carry out. This entails that recourse should be *personalized*, because different users in the same situation may *need* substantially different recourse plans. To see this, consider a user who is denied a loan. Based on a profile made by the financial institution, an AR algorithm might suggest the user reduce their monthly expenses. However, unlike the "average" customer, our user is incurring high medical expenses because they recently contracted an invalidating illness. Thus, the AR suggestion is highly inappropriate. Clearly, it is impossible to infer such a constraint from their profile alone. Most approaches, however, completely neglect the user's own preferences. The few that do require feedback that is difficult to obtain in practice, *e.g.*preferences over a large pool of alternatives [120, 121, 122, 123] or upfront quantification of action costs [124, 114, 125, 122, 126].

We argue that algorithmic recourse should make users first-class citizens in the recourse generation process rather than viewing them as passive observers. To this end, we introduce `PEAR` (Preference Elicitation for Algorithmic Recourse), the first human-in-the-loop approach for generating *personalized* recourse tailored for a target end-user. Our algorithm integrates AR and ideas from interactive Preference Elicitation (PE) [127, 128, 129, 130, 131, 132] in a fully Bayesian setup. `PEAR` goes beyond existing approaches in that the costs of actions are estimated from user feedback and prior information. In each iteration, `PEAR` identifies a *small* selection of alternative interventions – a *choice set* – that optimizes a sound measure of information gain (the *Expected Utility of Selection* (EUS) [131, 133]) and then asks the user to pick their preferred option. Using this feedback, `PEAR` quickly improves its estimate of the user's preferences and generates interventions that get progressively closer to the user's ideal. Furthermore, `PEAR` takes interactions and dependencies between actions costs into account when computing suggested recourse, whenever this information is available. See Fig. 16 for an overview.

Contributions. Summarizing, we:

- Introduce the problem of *personalized algorithmic recourse*, and show that existing approaches are insufficient to solve it.
- Develop `PEAR`, the first human-in-the-loop Bayesian approach for computing *personalized* interventions that is robust to noise in user feedback and minimizes user effort.
- Evaluate `PEAR` on synthetic and real-world datasets and show that it can generate substantially – up to 50% – cheaper interventions than user-agnostic competitors after only a handful of queries.

Problem Statement. The user *state* $\boldsymbol{s} = (s_1, \ldots, s_d) \in \mathcal{S} \subseteq \mathbb{R}^d$ is a vector of d categorical and real-valued features encoding, *e.g.*instruction level and income. An *action* $a \in \mathcal{A}$ is a map that takes a state $\boldsymbol{s}$ and changes a *single* feature, yielding a new state $\boldsymbol{s}' = a(\boldsymbol{s})$, and expresses a recommendation of the form "*Increase your income by $100*". Given a (black-box) binary classifier[5] $h : \mathcal{S} \to \{0, 1\}$ and a user state $\boldsymbol{s}$ leading to an undesirable decision $y = h(\boldsymbol{s})$, AR computes an *intervention* – *i.e.*an ordered set of actions $I = \{a^{(1)}, \ldots, a^{(|I|)}\}$ – that can be applied to $\boldsymbol{s}$ to obtain a counterfactual state $\boldsymbol{s}'$ associated to a more desirable outcome $h(\boldsymbol{s}') \neq y$, *all while minimizing user effort*. We formalize the user effort required to perform an action via a *cost function* $C : \mathcal{A} \times \mathcal{S} \to \mathbb{R}^+$. The cost of an intervention is the sum of the costs of all actions it contains, that is $C(I) = \sum_{i=0}^{|I|} C(a^{(i)}, \boldsymbol{s}^{(i)})$. Here, $\boldsymbol{s}^{(i)}$ is the state obtained by applying action

[5] It is straightforward to adapt our approach to deal with multiclass classification problems.

$a^{(i-1)}$ to state $\boldsymbol{s}^{(i-1)}$, and $\boldsymbol{s}^{(0)} = \boldsymbol{s}$ is the initial state. We denote with $I(\boldsymbol{s}^{(0)})$ the operation of applying each action $a \in I$ sequentially.

Approaches to AR assume the user effort is proportional to the number of actions that need to be carried out, and minimize it by searching for *short* interventions I. However, this is unrealistic and impractical. For instance, changing job into a highly skilled one may not be realistic without obtaining a Master's degree first. In reality, the user effort also depends on *users' preferences* $\boldsymbol{w} \in \mathcal{W} \subseteq \mathbb{R}^d$ which we represent as d-dimensional vectors. For example, for some people, it might be easier to improve their education than get new employment. Thus, we *parameterize* the cost function using the user preferences as $C(I \mid \boldsymbol{w})$, or $C(a, \boldsymbol{s} \mid \boldsymbol{w})$ if we refer to single actions. In the following, we distinguish between the (typically unknown) user preferences $\boldsymbol{w}^{GT}$ and the preferences $\boldsymbol{w}$ used by the recourse algorithm, which might differ from the former. Motivated by this, we introduce a new problem setting, denoted *personalized algorithmic recourse*:

Definition 1 (Personalized Algorithmic Recourse,) *Given a black-box binary classifier h and a user state $\boldsymbol{s}$, acquire a cost function $C(I \mid \boldsymbol{w})$ for interventions such that the intervention I^* obtained by solving the following optimization problem:*

$$I^* \in \arg\min_I \; C(I \mid \boldsymbol{w}) \quad \text{s.t.} \quad h(I(\boldsymbol{s}^{(0)})) \neq h(\boldsymbol{s}^{(0)}) \tag{14}$$

*has **minimal regret** for the target user, defined as:*

$$Reg(I^*, I^{GT}) = C(I^* \mid \boldsymbol{w}^{GT}) - C(I^{GT} \mid \boldsymbol{w}^{GT})$$

where $\boldsymbol{w}^{GT} \in \mathcal{W}$ encodes the ground-truth but unobservable preferences of the user and I^{GT} is the "ideal" intervention that would be obtained by solving Eq. (14) using $\boldsymbol{w}^{GT}$.

The similarity of Eq. (14) to existing formulations of AR can be misleading, as here the key challenge is that of obtaining weights $\boldsymbol{w}$ that reflect the user's own preferences.

3.5.2 The PEAR Algorithm

In order to account for uncertainty over the user's weights $\boldsymbol{w}$, PEAR explicitly models a distribution $P(\boldsymbol{w})$ over them and progressively refines it by interacting with a target user. A high-level overview of PEAR is given in Fig. 16 and the pseudo-code is listed in Algorithm 1.

In each iteration $t = 1, \ldots, T$, where T is the iteration budget, PEAR computes a *choice set* $O^{(t)} \in \mathcal{I}^k$ containing k candidate interventions achieving recourse (for a small k, *e.g.*2 to 4) and asks the user to indicate their most preferred option in the set. Importantly, $O^{(t)}$ is chosen so as to maximize the (expected) information gained from the user, and in a way that is robust to noise in their feedback. We detail the exact procedure used by PEAR in Section 3.5.2. These user choices are stored in an initially empty dataset $\mathcal{D}^{(t)}$. In each step, PEAR integrates the user's feedback by inferring a posterior over the weights $P(\boldsymbol{w} \mid \mathcal{D}^{(t)}) \propto P(\mathcal{D}^{(t)} \mid \boldsymbol{w})P(\boldsymbol{w})$ using Bayesian inference, and updates the user state by applying the first action $\hat{I}_1$ of the chosen intervention. We apply a single action so as to elicit user preferences in all intermediate states. If a state achieving recourse is reached, the user state is reinitialized. After T rounds,[6] PEAR computes a low-cost personalized intervention by applying the intervention generation procedure described in Section 3.5.2, biased according to the latest posterior $p(\boldsymbol{w} \mid \mathcal{D}^{(t)})$.

PEAR makes no assumption on the form of the prior $P(\boldsymbol{w})$, meaning that the prior can be adjusted based on the application. In order to model both variances across the preferences of individuals and for sub-groups in the population, in this work we model them as a mixture of Gaussians

[6] In practice, the loop can be terminated as soon as the user is satisfied with one of the interventions in $O^{(t)}$.

Algorithm 1 The PEAR algorithm: $h : \mathcal{S} \to \{0, 1\}$ is a classifier, $\boldsymbol{s}^{(0)} \in \mathcal{S}$ the initial state, $\mathcal{A}$ the available actions, $p(\boldsymbol{w})$ the prior, $T \geq 1$ the query budget, $k \geq 2$ is the size of choice sets.

procedure PEAR($h, \boldsymbol{s}^{(0)}, \mathcal{A}, T, k$)
  Initialize $t \leftarrow 0$, $\mathcal{D}^{(0)} \leftarrow \emptyset$
  for $t = 1, \ldots, T$ **do**
    $O^{(t)} \leftarrow$ SUBMOD-CHOICE($h, \boldsymbol{s}^{(t-1)}, \mathcal{A}, k, \mathcal{D}^{(t-1)}$) ▷ Algorithm 2
    Ask the user to pick the best intervention $\hat{I} \in O^{(t)}$
    $\mathcal{D}^{(t)} \leftarrow \mathcal{D}^{(t-1)} \cup \{\hat{I}\}$
    Update weight estimate $p(\boldsymbol{w} \mid \mathcal{D}^{(t)})$
    $\boldsymbol{s}^{(t)} \leftarrow \hat{I}_1(\boldsymbol{s}^{(t-1)})$
    if $h(\boldsymbol{s}^{(t)}) \neq h(\boldsymbol{s}^{(0)})$ **then**
      $\boldsymbol{s}^{(t)} \leftarrow \boldsymbol{s}^{(0)}$
  $I^* =$ W-FARE($h, \boldsymbol{s}^{(0)}, \boldsymbol{w}^*$) with $\boldsymbol{w}^* = \mathbb{E}_{P(\boldsymbol{w} \mid \mathcal{D}^{(T)})}[\boldsymbol{w}]$ ▷ Section 3.5.2
  return I^*

with M components $\mathcal{N}_i(\mu_i, \ _i)$, for $i = 1, \ldots, M$. This choice works well in our experiments, see Section 3.5.3. Note that, analogously to [125], it is also possible to fit the prior on population-level preference data or domain expert input, whenever this is available. In our experiments, we do this for *all* competitors.

Generating Personalized Interventions with W-FARE. PEAR generates personalized interventions by leveraging a novel, *user-aware* extension of FARE [112], a state-of-the-art algorithm for generating short – but *user-agnostic* – interventions, which we briefly outline next. In FARE, each action $a \in \mathcal{A}$ is implemented as a tuple (f, x), where f is a function changing *one* feature and x is the value that feature takes, *e.g.*(*change_income*, \$1000). Given an initial state $\boldsymbol{s}$, FARE uses reinforcement learning to learn two probabilistic policies $\pi_f(\boldsymbol{s})$ and $\pi_x(\boldsymbol{s})$, which are used as priors to guide a Monte Carlo Tree Search procedure that incrementally builds an intervention I by selecting actions $a^{(i)} \in \mathcal{A}$. In order to ensure interventions are *actionable*, actions a are only chosen if they satisfy given preconditions. The reward used by FARE is $r(I) = \rho^{|I|} \cdot \mathbb{1}\{h(I(\boldsymbol{s}^{(0)})) \neq h(\boldsymbol{s}^{(0)})\}$, where $\rho > 0$ is a discount factor and the indicator evaluates to 1 if I attains recourse and to 0 otherwise. FARE is highly scalable and very effective at identifying counterfactual interventions even under minimal training budget [112].

FARE is user-agnostic, while PEAR needs to generate *personalized* interventions. We fill this gap by introducing W-FARE, a novel extension of FARE that integrates the user's costs into the reward while inheriting all benefits of the latter. Recall that PEAR maintains a posterior over the weights. The *expected cost of an action* can thus be obtained by marginalizing over the posterior:

$$\mathbb{E}[C(a, \boldsymbol{s}) \mid \mathcal{D}^{(t)}] = \int_{\boldsymbol{w}} C(a, \boldsymbol{s} \mid \boldsymbol{w}) P(\boldsymbol{w} \mid \mathcal{D}^{(t)}) \, d\boldsymbol{w}$$

Analogously, the cost of an intervention I is replaced by the expectation:

$$\mathbb{E}[C(I) \mid \mathcal{D}^{(t)}] = \sum_{i=0}^{|I|} \mathbb{E}[C(a^{(i)}, \boldsymbol{s}^{(i)}) \mid \mathcal{D}^{(t)}]$$

The W-FARE reward function is then given by $r(I \mid \boldsymbol{w}) \propto \rho^{\mathbb{E}[C(I \mid \mathcal{D}^{(t)})]} \cdot \mathbb{1}\{h(I(\boldsymbol{s}^{(0)})) \neq h(\boldsymbol{s}^{(0)})\}$. This explicitly drives RL to learn policies that optimize user-specific action costs and that, therefore, help MCTS to more quickly converge to *personalized* interventions. We show empirically that PEAR is substantially more effective than FARE at computing personalized interventions in Section 3.5.3.

Computing Informative Choice Sets. Given the current posterior $P(\boldsymbol{w} \mid \mathcal{D}^{(t)})$, PEAR computes a *choice set* containing k interventions I that maximizes information gain [127]. We measure the latter using the *Expected Utility of Selection* (EUS) [133], a measure of the goodness of a set defined as the expectation, under the uncertainty over $\boldsymbol{w}$, of the utility of its most preferred element. EUS is closely related to the Expected Value of Information (EVOI), and frequently used in Bayesian PE [134, 131, 135, 133]. The EUS builds on the notion of *expected utility of an intervention* I, which is defined as:

$$\mathsf{EU}(I \mid \mathcal{D}^{(t)}) = \mathbb{E}[-C(I) \mid \mathcal{D}^{(t)}] = -\int_{\boldsymbol{w}} C(I \mid \boldsymbol{w}) P(\boldsymbol{w} \mid \mathcal{D}^{(t)})\, d\boldsymbol{w}$$

The EUS of a choice set O can then be defined as:

$$\begin{aligned} \mathsf{EUS}_R(O \mid \mathcal{D}^{(t)}) &= \textstyle\sum_{I \in O} P_R(O \rightsquigarrow I) \mathsf{EU}(I \mid \mathcal{D}^{(t)}) \\ &= -\int_{\boldsymbol{w}} \left[\textstyle\sum_{I \in O} P_R(O \rightsquigarrow I \mid \boldsymbol{w}) C(I \mid \boldsymbol{w})\right] P(\boldsymbol{w} \mid \mathcal{D}^{(t)})\, d\boldsymbol{w} \end{aligned} \tag{15}$$

Here, $P_R(O \rightsquigarrow I \mid \boldsymbol{w})$ is the probability that a user with weights $\boldsymbol{w}$ picks I from O, under a specific choice of *response model* R modelling noise in user choices. Intuitively, we expect users to prefer the cheapest interventions $I \in O$. Moreover, we also expect that interventions in O with similar costs have a similar probability of being chosen. Motivated by this, and following common practice in choice modelling [136], in PEAR we implement a *logistic response model* (L), defined as:

$$P_L(O \rightsquigarrow I \mid \boldsymbol{w}) = \frac{\exp(-\lambda C(I \mid \boldsymbol{w}))}{\sum_{I \in O} \exp(-\lambda C(I \mid \boldsymbol{w}))}$$

Here, $\lambda \in \mathbb{R}$ is a temperature parameter. Finding a choice set O maximizing the EUS is intractable in general – *NP-hard* [137, 138], in fact – and computationally intensive in practice, and risks slowing down the interaction loop to the point of estranging users. We observe that, however, under some response models R, the EUS becomes *submodular* and *monotonic* [131]. This is the case for the *noiseless* response model (NL), according to which the user always prefers the lowest-cost option, *i.e.*[7]

$$P_{NL}(O \rightsquigarrow I \mid \boldsymbol{w}) = \textstyle\prod_{I, I' \in O \,:\, I \neq I'} \mathbb{1}\{C(I \mid \boldsymbol{w}) < C(I' \mid \boldsymbol{w})\}$$

This means that, for NL, greedy optimization is sufficient to find a choice set O that achieves high EUS_{NL} with approximation guarantees. Formally, it holds that for choice sets O found via greedy optimization, $\mathsf{EUS}_{NL}(O \mid \mathcal{D}) \geq (1 - e^{-1})\mathsf{EUS}_{NL}(O^* \mid \mathcal{D})$, where O^* is the truly optimal choice set [131, 137, 138]. In PEAR, we leverage the fact that $\mathsf{EUS}_L - \mathsf{EUS}_{NL}$ is always smaller than a problem independent (tight) bound [131], meaning that instead of minimizing EUS_L directly, we can compute a high-quality choice set by greedily maximizing EUS_{NL}. This immediately leads to a practical algorithm for the logistic response model L, listed in Algorithm 2.

3.5.3 Experiments and Results

Our experimental evaluation is aimed at answering the following research questions:

Q1 Does PEAR succeed in minimizing regret for an increasing amount of user feedback?

Q2 Does PEAR outperform competitors in terms of validity and cost?

Q3 Is PEAR robust to imprecise knowledge of the cost correlation structure of the user?

[7] For the sake of presentation, we assume that there are no ties. Note that the EUS formula is invariant to the way ties are broken. In our implementation, ties are broken uniformly at random.

Algorithm 2 Greedy procedure to efficiently compute a choice set O: $\boldsymbol{s}^{(t)} \in \mathcal{S}$ the current state, $\mathcal{A}$ the available actions, $k \geq 2$ is the size of choice sets, $\mathcal{D}^{(t)}$ the user choices so far.

procedure SUBMOD-CHOICE($\boldsymbol{s}^{(t)}, k, \mathcal{A}, \mathcal{D}^{(t)}$)
  $O \leftarrow \emptyset$
  $\bar{\boldsymbol{w}} \leftarrow \mathbb{E}_{p(\boldsymbol{w}|\mathcal{D}^{(t)})}[\boldsymbol{w}]$
  while $|O| < k$ **do**
    Generate the candidate interventions $\mathcal{I}$ with W-FARE using $\mathcal{A}$ and $\bar{\boldsymbol{w}}$
    $\hat{I} \leftarrow \text{argmax}_{\hat{I}}\ \mathsf{EUS}_{NL}(O \cup \hat{I} \mid \mathcal{D}^{(t)}) - \mathsf{EUS}_{NL}(O \mid \mathcal{D}^{(t)})$
    $O \leftarrow O \cup \{\hat{I}\}$
  return O

Datasets and Classifiers. We evaluated our approach on two real-world datasets taken from the relevant literature: `GiveMeSomeCredit` [139] and `Adult` [140]. They are (unbalanced) binary classification problems for income prediction and loan assignment, respectively. The datasets have both categorical and numerical features. Some of these features are actionable (e.g., *occupation*, *education*), while others are immutable (e.g., *age*, *sex*, *native_country*). These datasets come without a causal graph and users' preferences over the features. Following previous work [111, 112, 113], we manually defined and fixed the cost correlation structures for both. We randomly generated user-specific weights for each instance by sampling from a (dataset-specific) mixture of Gaussians with $M = 6$ components and uniform mixture weights. We then split the data into training (70%), validation (10%) and test (20%) sets. For each dataset, we designed the black-box classifier h as an MLP with two hidden layers. We trained it by cross-entropy minimization, selecting the hyperparameters which maximise the F_1 score on the validation set.

"Easy" vs. "Hard" Users. Intuitively, users close to the decision boundary of the black-box h will require few actions to achieve recourse, while users to whom h assigns a low score might need longer and more complex interventions. Understanding the preferences of these "hard" users is crucial since a wrong suggestion might substantially increase the overall cost for them. For each dataset, we thus built two separate testing sets. The first one, named `All`, is obtained by sampling 300 users $\boldsymbol{s}$ with an unfavourable classification ($h(\boldsymbol{s}) < 0.5$), regardless of the actual value of h. The second one, named `Hard`, is obtained by sampling 300 users with an unfavourable classification having a score in the lower quartile of the black-box score distribution.

Competitors. We compare `PEAR` against several baselines: `FARE` and its explainable version `EFARE` [112], `CSCF` [113], an evolutionary algorithm which, similarly to `FARE`, generates recourse options by considering consequence-aware cost functions and action sets $\mathcal{A}$, and `FACE` [141], a well-know AR algorithm, which optimizes for population-based "feasible paths" to achieve recourse. We also consider two simpler baselines, a brute-force search (`MCTS`) and a vanilla reinforcement learning agent (`RL`), trained in a similar way as in [142, 143]. Note that all the competitors are model-agnostic and *not* interactive, since they assume the users' costs to be fixed.

Experimental Protocol. For `PEAR`, we vary the number of questions T to the user from 0 to 10. For $T = 0$, we initialize the weights with the expected value of the prior, $\mathbb{E}_{P(\mathbf{w})}[\mathbf{w}]$, that represents a user-independent population-based prior. Moreover, we employ two user response models, the *noiseless* model (Section 3.5.2), to check the effectiveness of our approach in the best-case scenario where the user can perfectly express their preferences, and the *logistic* model (Section 3.5.2), to

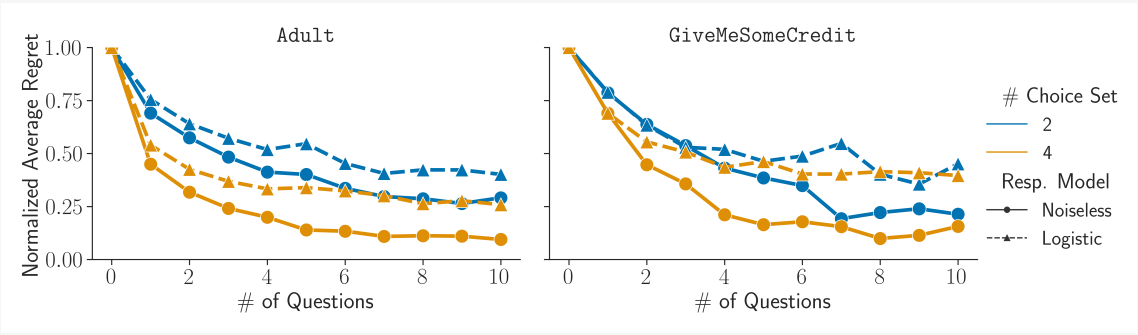


Figure 17. Normalized Average Regret for `PEAR` *when varying the number of questions, the choice set size and the user response model on both datasets (sampled from* `All` *users).*

challenge our approach in a more realistic scenario. To provide a fair comparison, we equip the competitors with our cost function and set their weights to the expected value of the prior.

Q1: `PEAR` **Successfully Minimizes the Regret.** Fig. 17 shows the evaluation of the regret as a function of the number of queries to the user. Here the ground-truth intervention I^{GT} (which is unknown) is approximated by running `PEAR` with the correct user costs $\boldsymbol{w}^{GT}$, and the regret is normalized by rescaling the costs between $C(I^{\text{GT}} \mid \boldsymbol{w}^{GT})$ and $C(I^{(0)} \mid \boldsymbol{w}^{GT})$ where we generate $I^{(0)}$ using the expectation of the prior. We run `PEAR` with two different choice set dimensions, $k = 2$ and $k = 4$, and for both noiseless and logistic response models. After a few questions, `PEAR` reaches a low regret in all settings. Generally, a larger choice set produces a lower regret, irrespective of the response model, with the downside of increasing the cognitive burden for the user. We now briefly summarize the results when $T = 10$. For the `Adult` dataset, the best regret is ≈ 0.09 for the noiseless user and $k = 4$, while the worst regret is ≈ 0.40 for the logistic response model and $k = 2$. For `GiveMeSomeCredit`, we get ≈ 0.15 (noiseless, $k = 4$) and ≈ 0.45 (logistic, $k = 2$). Overall, we can provide interventions which are at least 50% cheaper than their preference-agnostic counterparts.

Q2: `PEAR` **Outperforms Competitors in terms of Validity and Cost.** Following the AR literature [114, 144], we compare `PEAR` (with $T = 10$) and all competitors in terms of average *validity*, *i.e.*fraction of users for which we obtained recourse, intervention *cost* and *length* (or *sparsity*), *i.e.*the number of features that have to be changed. Intervention costs are computed by using the true weights $\boldsymbol{w}^{GT}$. Table 8 shows the results. `PEAR` manages to achieve the highest validity while also providing substantially cheaper interventions than the non-personalized competitors on average. This is true both for the noiseless and logistic response models. While `CSCF` tends to produce shorter interventions, these are in general more costly and have a larger cost variance with respect to those found by `PEAR`, confirming the intuition that length is a suboptimal proxy of intervention complexity. The only exception is the `Hard` setting of `GiveMeSomeCredit`, where however `CSCF` manages to achieve recourse for only 22% of the users, whereas `PEAR` achieves recourse in 58% of the cases. The difficulty of `CSCF` in achieving recourse is visible in all settings and severely limits its applicability. Furthermore, `CSCF` is 10 to 50 times more computationally expensive than `PEAR`, making it unsuitable for real-time interactive scenarios. The `MCTS` baseline has rather poor performance both in terms of validity and cost in all settings, while the lbaseline has a reasonably high validity on `Adult` but it completely fails to learn a policy achieving recourse on `GiveMeSomeCredit`. On the other hand, methods which combine `MCTS` and l(`FARE` and `EFARE`)

Users	Method	Adult			GiveMeSomeCredit		
		Validity	*Cost*	*Length*	*Validity*	*Cost*	*Length*
All	FARE	0.90 ± 0.27	285.63 ± 195.68	3.45 ± 1.19	0.86 ± 0.23	161.77 ± 107.42	3.27 ± 1.13
	CSCF	0.78 ± 0.28	154.28 ± 125.34	2.53 ± 0.57	0.57 ± 0.42	100.69 ± 120.22	2.51 ± 1.12
	EFARE	0.76 ± 0.39	306.11 ± 199.02	3.54 ± 1.25	0.67 ± 0.38	155.92 ± 109.19	3.18 ± 1.18
	RL	0.76 ± 0.38	283.31 ± 167.86	3.36 ± 1.08	0.12 ± 0.32	59.66 ± 0.00	2.00 ± 0.00
	MCTS	0.44 ± 0.44	445.65 ± 201.28	4.60 ± 1.19	0.70 ± 0.42	214.33 ± 119.38	4.09 ± 1.31
	FACE	0.15 ± 0.27	397.49 ± 128.43	3.76 ± 0.64	0.24 ± 0.38	327.18 ± 78.85	5.97 ± 0.62
	PEAR_{NL} **(ours)**	$\mathbf{1.00 \pm 0.03}$	$\mathbf{142.23 \pm 61.75}$	$\mathbf{2.84 \pm 0.59}$	$\mathbf{0.89 \pm 0.00}$	$\mathbf{96.04 \pm 31.96}$	$\mathbf{2.79 \pm 0.42}$
	PEAR_{L} **(ours)**	$\mathbf{1.00 \pm 0.04}$	$\mathbf{146.54 \pm 63.09}$	$\mathbf{2.84 \pm 0.56}$	$\mathbf{0.89 \pm 0.00}$	$\mathbf{100.19 \pm 29.53}$	$\mathbf{2.85 \pm 0.48}$
Hard	FARE	0.71 ± 0.44	438.97 ± 188.50	4.54 ± 1.21	0.47 ± 0.37	319.46 ± 96.36	4.97 ± 0.70
	EFARE	0.55 ± 0.48	454.05 ± 202.76	4.52 ± 1.25	0.22 ± 0.36	371.58 ± 82.18	5.31 ± 0.71
	RL	0.55 ± 0.47	433.64 ± 152.67	4.33 ± 1.24	-	-	-
	CSCF	0.25 ± 0.36	382.84 ± 126.31	3.70 ± 0.56	0.13 ± 0.32	190.81 ± 119.99	3.36 ± 1.11
	MCTS	0.21 ± 0.40	599.20 ± 153.44	5.53 ± 0.72	0.40 ± 0.43	353.75 ± 99.26	5.43 ± 0.88
	FACE	0.00 ± 0.04	448.72 ± 0.00	5.20 ± 0.00	0.20 ± 0.36	455.20 ± 92.10	7.09 ± 0.53
	PEAR_{NL} **(ours)**	$\mathbf{0.99 \pm 0.08}$	$\mathbf{296.37 \pm 43.84}$	$\mathbf{3.35 \pm 0.55}$	$\mathbf{0.58 \pm 0.04}$	$\mathbf{251.60 \pm 51.16}$	$\mathbf{4.59 \pm 0.40}$
	PEAR_{L} **(ours)**	$\mathbf{0.99 \pm 0.09}$	$\mathbf{301.13 \pm 52.61}$	$\mathbf{3.34 \pm 0.58}$	$\mathbf{0.58 \pm 0.02}$	$\mathbf{262.23 \pm 45.36}$	$\mathbf{4.64 \pm 0.36}$

Table 8. Performance of all competitors averaged over 10 runs. A '-' indicates that the method did not find any successful intervention for any user. PEAR_{NL} *and* PEAR_{L} *indicate PEAR associated with the noiseless and logistic response model, respectively. The best results are boldfaced.*

give better performance, which is aligned with previous results [112], but are still suboptimal with respect to PEAR in terms of both validity and cost. Finally, FACE struggles to achieve recourse since it needs to find a "feasible path" from the current user to a similar one *in the training set*, which is favourably classified.

Q3: PEAR is Robust to Misspecifications of the cost Correlation Structure. In the previous experiments, following other research works [111, 112, 113], we assumed to know the structure of the CCS a-priori. However, in a real scenario, we might have instead an *approximate* causal graph from which to derive the CCS. Table 9 shows the validity, cost and length of the interventions found by removing $X\%$ of edges from the causal graph, with $X \in \{0.15, 0.25, 0.50, 1.00\}$. Validity is almost unaffected by corruption in all settings since it only impacts the computation of the cost. On the other hand, as expected, increasing the amount of graph corruption reduces the effectiveness of user feedback. However, the degradation is not dramatic. Indeed, if we look at the Hard evaluation, the increase in cost is negligible (around 4%) with up to 50% randomly removed edges. On GiveMeSomeCredit, we do not see any significant increase in costs. Surprisingly, we see instead an improvement for 15% and 25% corruption levels. We hypothesize that lacking causal knowledge about features which are not needed for recourse can be beneficial since it simplifies the elicitation process. However, at higher levels of corruption, this effect disappears. The setting $X = 1.0$ is equivalent to a non-causal cost function, in which acting on a feature has always the same cost, irrespective of the others. It is the common choice of many works dealing with AR [145, 115, 141]. Under such a setting, when considering All users, the degradation is more evident, but still within 6% for GiveMeSomeCredit, while for Adult it goes up to 15%. Overall, results clearly indicate that PEAR can suggest reasonable cost interventions even with a largely misspecified cost correlation structure. This is apparent when comparing these results with those in Table 8. Even with 50% randomly removed edges, PEAR recommends interventions that are cheaper than

Users	Corruption	Adult			GiveMeSomeCredit		
		Validity	*Cost*	*Length*	*Validity*	*Cost*	*Length*
All	**None**	1.00 ± 0.04	146.54 ± 63.09	2.84 ± 0.56	0.89 ± 0.00	100.19 ± 29.53	2.85 ± 0.48
	0.15	1.00 ± 0.00	165.11 ± 46.19	2.79 ± 0.24	0.90 ± 0.00	98.02 ± 18.37	2.46 ± 0.16
	0.25	1.00 ± 0.00	162.05 ± 48.05	2.89 ± 0.34	0.90 ± 0.07	99.51 ± 18.53	2.48 ± 0.18
	0.5	1.00 ± 0.00	175.54 ± 62.29	2.82 ± 0.36	0.90 ± 0.11	110.35 ± 19.93	2.56 ± 0.20
	1.0	0.99 ± 0.02	172.51 ± 61.13	2.96 ± 0.40	0.91 ± 0.04	106.63 ± 28.84	2.64 ± 0.28
Hard	**None**	0.99 ± 0.09	301.13 ± 52.61	3.34 ± 0.58	0.58 ± 0.02	262.23 ± 45.36	4.64 ± 0.36
	0.15	1.00 ± 0.00	308.22 ± 38.40	3.47 ± 0.26	0.63 ± 0.04	237.00 ± 35.72	4.06 ± 0.31
	0.25	1.00 ± 0.02	305.62 ± 39.73	3.32 ± 0.35	0.66 ± 0.11	238.92 ± 28.66	3.90 ± 0.27
	0.5	1.00 ± 0.03	314.36 ± 37.93	3.38 ± 0.32	0.59 ± 0.14	250.47 ± 29.71	4.16 ± 0.24
	1.0	0.99 ± 0.03	313.51 ± 61.64	3.69 ± 0.44	0.64 ± 0.09	256.24 ± 12.04	4.24 ± 0.18

Table 9. Evaluation of `PEAR` *(with $q = 10$ and a logistic noise model) for an increasing amount of CCS graph corruption, averaged over 10 runs. "None" indicates that the correct causal graph is being used.*

all competitors but `CSCF`, that however has a substantially lower validity.

3.5.4 Conclusion

In this work, we identify the problem of *personalized* algorithmic recourse as a fundamental stepping stone for ensuring recourse is usable in real-world applications, and develop `PEAR`, the first algorithm able to provide *personalized* interventions. Our experimental evaluation shows that `PEAR` substantially outperforms existing (non-personalized) solutions in terms of both validity and intervention cost with only a handful of queries to the user. We hope that this initial contribution can foster further research in the community to work towards a more realistic form of algorithmic recourse that can be successfully deployed in real-world scenarios. As for all methods dealing with algorithmic recourse, the effectiveness of the approach should, in principle, be evaluated on real users. However, this evaluation is highly non-trivial (and thus still missing in the algorithmic recourse literature) because it requires the creation of a realistic scenario where a user feels to be *unfairly treated* in some machine-driven decision involving their life. The legal requirements that are progressively being introduced to regulate AI systems [146] could contribute to making the information needed to set up such a scenario available in the near future.

3.5.5 Relevant Publications

- [147] G. D. Toni, P. Viappiani, S. Teso, B. Lepri, and A. Passerini, "Personalized algorithmic recourse with preference elicitation," *Transactions on Machine Learning Research*, 2024, https://openreview.net/forum?id=8sg2I9zXgO
- [148] S. Esfahani, G. De Toni, B. Lepri, A. Passerini, K. Tentori, and M. Zancanaro, "Preference elicitation in interactive and user-centered algorithmic recourse: an initial exploration," in *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, 2024, https://dl.acm.org/doi/pdf/10.1145/3627043.3659556

3.5.6 Relevant Software and/or External Resources

- https://github.com/unitn-sml/pear-personalized-algorithmic-recourse

3.6 Quantifying Fairness with Fuzzy Logic

Contributing partners: CERTH

3.6.1 Introduction and Methodology

Quantifying bias and fairness concerns for AI systems with the goal of creating fairer ones is an emerging subject of many research papers [149, 53, 150] and algorithmic frameworks [151] like AIF360 [152] and FairLearn [153]. However, measures that perform such quantification tend to employ ad-hoc definitions of what is "fair" without clarifying which policymaker and stakeholder opinions they reflect. Involving the opinions of people outside the research circles creates socially responsible AI [154] by enabling participatory design and democratic deliberation [155]. Conversely, without explicit acknowledgement of which concerns are quantified by measures, it is unclear how to port them in real-world contexts different than those they were originally conceived for. Furthermore, multiple measures may be mathematically or conceptually incompatible with each other [156], and only those suited to the context should be selected.

To address the issue of making algorithmic fairness research outcomes practically useful, we look at a framework [157] that disentangles formal definitions of fairness from context-specific choices (e.g., that determine hyperparameter values) and investigate how it can be adopted by real people to create definitions of bias or fairness. This framework uses fuzzy logic [158], which is broader than classical logic, being thus able to reflect uncertain opinions. The framework consists of three steps: A) Bias or fairness definitions are dictated by policymakers and transcribed to basic fuzzy logic [159, 160] statements that use abstract predicates and logical connectives (and, or, etc.). B) Context-specific predicate understanding and truth values are extracted from stakeholder beliefs. C) basic fuzzy logic evaluation rules [161, 162, 163] yield the corresponding truth value of the definitions.

Practical adoption of the framework requires a collaboration between actors across multiple disciplines that bring different types of expertise. Thus, in Subsection 3.6.2, we identify these actors and create an interdisciplinary process to help them both collaborate and combine policymaker and stakeholder opinions. In Subsection 3.6.3, we then recognize that one of the key practical challenges in creating context-specific definitions of fairness lies with the repetition of costly and time-consuming stakeholder opinion gathering processes. For this reason, we present methods for simplifying these processes.

3.6.2 Interdisciplinary Collaboration

An interdisciplinarity collaboration is needed for the practically grounded adoption of the fuzzy logic framework described above. Here, we summarize the involved actors and their collaboration workflow.

Actors. First, fuzzy *logic practitioners* or other researchers with the required expertise should accurately express in basic fuzzy logic the context-independent part of fairness definitions that reflect the intent of policymakers, be they common assumptions or regulation. Note that familiarization with this kind of logic can be arduous without prior experience and, for this reason, processes similar to causal model building [164, 165] should be employed. If the end-result of logical analysis is standardized for a broad range of assumptions or legal settings, AI creators can determine the standardization's parameters based on domain expertise, and only the step of encoding stakeholder beliefs remains. Second, *social scientists* should gather stakeholder feedback with which to

quantify (obtain the truth value in the specific context) abstract predicates found in basic fuzzy logic definitions. In addition to gathering and parsing potentially conflicting viewpoints, social scientists should also be the ones to lead discussions on how predicates being quantified should be interpreted, and negotiate which criteria belief truth values should be obtained. As quantitative feedback may be hard to gather from lay people, the burden of a converting qualitative discussions to a numerical scale may fall on the social scientists gathering it, with accompanying subjectivity concerns. Questionnaires or other means of extracting stakeholder beliefs may also require interdisciplinary design, for example, to account for edge cases both algorithmically (see below) and socially. Third, *computer scientists* (e.g., data analysts, machine learning engineers) are needed to implement the extracted definitions of fairness or bias. These are primarily the AI creators that should drive the adoption of this work in practice.

Workflow. Fostering a collaboration between the above actors presents its own challenges, ranging from organizational complexity to reconciling different viewpoints (e.g., sociological vs technical). One prospective scheme that reduces the interactions between disciplines is presented in Figure 18. In this, logic experts create logical formulas, social scientists determine truth values of predicates, and computer scientists run the fuzzy logic framework. Importantly, these processes run independently, which means that the points of contact between different disciplines are minimized. There is only one feedback loop in which the opinions of stakeholers at large affect the abstract fairness definitions provided by policymakers. Computer scientists and logic experts could be the same people. Logic experts and social scientists gather context-independent and context-specific opinions, respectively encapsulated in abstract fairness definitions and in truth values.

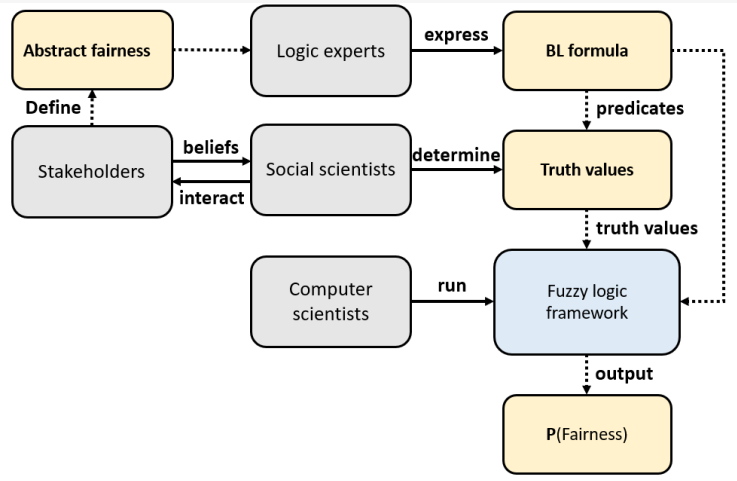


Figure 18. Interdisciplinary collaboration to apply our framework in practice.

3.6.3 Learning to Replicate Stakeholder Beliefs

Given that policymaker opinions see widespread adoption, it remains to quantify context-specific stakeholder beliefs. This quantification can often be simplified by gathering beliefs under multiple artificially constructed scenarios, and using these as a reference to extrapolate generic definitions of fairness. In the following we present theoretical details of the fuzzy logic framework, based on

which we make recommendations on how it can be used to extract fairness definitions in practice.

Preliminaries. An evaluation mechanism L of basic fuzzy logic is characterized by a t-norm operation $\star_L$ that computes the truth value of idempotent logical conjunction (two statements holding true simultaneously). The t-norm also defines a corresponding residuum operation $\Rightarrow_L$ that computes the truth value of logical implication. Different evalutation mechanisms are equivalent to additional axioms and corresponding logics; three base ones are demonstrated in Table 10, where the Mostert-Shields theorem [166] states that all other evaluation mechanisms are locally isomorphic to those.

Logic	$x \star_L y$	$x \Rightarrow_L y$
Gödel	$\min\{x, y\}$	$x\{1 \text{ if } x \leq y, y \text{ otherwise}\}$
Product	$x \cdot y$	$\{1 \text{ if } x \leq y, \frac{y}{x} \text{ otherwise}\}$
Lukasiewicz	$\min\{x + y - 1, 0\}$	$\min\{1, 1 - x + y\}$

Table 10. Three basic fuzzy logic evaluation mechanisms.

We present a theorem that is used in the rest of our analysis and covers definitions based on two main predicates: i) group membership, and ii) discrimination. The exact interpretation of these predicates is left to stakeholders. For example, one interpretation of group membership would be the chance of encountering protected group members, but could also indicate the chance of encountering a specific person in case of individual fairness.

THEOREM 2 OF [157]. *Let the predicate imbalance a) be an implication from group membership to some property, b) be certain for no group members and c) imply discrimination. If bias is imbalance for group members, then its truth value can be written as:*

$$bias = s \star_L e \star_L f(s, e) \leq s \star_L e$$

where s, e are the truth values of group membership and discrimination respectively, and $f(s, e)$ is a real-valued function for which $f(s, e) \leq f_{gen}(s, e) = \big(e \Rightarrow_L \inf_{s \star d \geq s \star e} d\big)$. The inequality is strict for $f(s, e) = f_{gen}(s, e)$. Fairness can be expressed as mutually exclusive to bias, which has truth value of $fairness = (bias \Rightarrow_L 0)$.

Intuitive Interpretation. When obtaining the truth value of bias or fairness from the above theorem, we argue that group membership will typically be a context-specific numerical constant, discrimination an (unobserved) function of the AI system and the data it is working with, and the function f a reflection of additional assumptions that underestimate the degree of bias. For example, if $f(s, e) = f_{gen}(s, e)$, which is the theoretical worst case for bias, $fairness = 0$ for all $s, e > 0$ in the Gödel and Product logics, additional assumptions are needed to even enable the prospect of achieving fairness with non-exact discrimination mitigation [167]. The same issue is not encountered in Lukasiewicz's logic, though.

We now explore two strategies for defining the truth value e of the discrimination predicate. The first strategy is common across the algorithmic fairness literature and depends on defining discrimination as the complement of some notion of indistinguishability between a group and the total population, such as fundamental differences in representation or treatment methods. In this case, stakeholders should be consulted to identify which types of differences matter. The second

strategy is more participatory and uses machine learning to directly learn to replicate stakeholder beliefs given an a-priory identification of relevant measurable system quantities.

Discrimination Through Indistinguishability. To define the discrimination predicate through further axiomatization, we rely on some base quantitative assessment $m \in [0, 1]$ for the group, and the same assessment $m' \in [0, 1]$ for some other group, such as the total population or the rest of the population. For example, the groups could correspond to a sensitive attribute's values, like Male vs Female in simplistic scenarios, the assessment could be representation, i.e., the fraction of positive samples, or misclassification rates. We now propose that the assement is the truth value of some property that should be indistinguishable between the groups. In basic logic terms, this means that the property holding for one group also implies that it holds for the other and conversely. We finally take the fuzzy set negation of the result to make the discrimination complementary to indistinguishibility with following formula:

$$e = 1 - (m \Rightarrow_L m') \star_L (m' \Rightarrow_L m)$$

In this equation, we substitute the evaluation mechanisms of some base logics to obtain respective truth values of discrimination in Table 11. From these, we observe that literature practices of comparing groups based on differences or ratios correspond to Lukasiewicz and Product logic respectively. We also obtain a justification of worst-case performance (e.g., worst-case accuracy) as a type of Gödel logic discrimination. The logic evaluation mechanism may be a direct result of stakeholder feedback based on how truth values degrade when multiple inexact truths are involved (e.g., in Lukasiewicz an a large enough number of non-exact statements will eventually yield 0 truth value under indempondent conjunction). Therefore, our analysis maps literature practices to intuitive terms to help select the mechanism to compute the truth value of discrimination from common options.

Logic	e
Gödel	$1 - \min\{m, m'\}$
Product	$1 - \frac{\min\{m,m'\}}{\max\{m,m'\}}$
Lukasiewicz	$\lvert m - m' \rvert$

Table 11. Truth values of discrimination e under basic fuzzy logic evaluation mechanisms. For $m = m' = 0$ Product logic discrimination evaluates to 0.

Learnable Discrimination. A simple methodology for extracting context-specific stakeholder beliefs would consist of gathering these beliefs through questionnaires. We thus introduce a process that constructs a broad enough range of synthetic edge cases and then gathers stakeholder opinions with social science processes, such as questionnaires. Then, consulting similar stakeholders can be replaced by interpolating between gathered opinions using deep learning. Here we show a full case study on this practice; this consists of an example fairness definition and a synthetically generated collection of stakeholder answers and negotiation outcomes.

Preliminary: Fairness Definition
In general, definitions of fairness will not be too complicated, as we want them to be intuitive and highly interpretable. If there are no other propositions to be modelled, worst-case fairness obtained for $f(s, e) = f_{gen}(s, e)$ should suffice. We expect more complex forms to arise when there are several

definitions to be simultaneously assessed, or when there are complex propositions to conjunct. For the sake of demonstration, let us assume a "business necessity" predicate with truth value b that, when it occurs for group members, should never imply discrimination, i.e., $((b \star_L s) \Rightarrow_L e) \Rightarrow 0$. This could be part of affirmative action that hires group members at a higher position, though keep in mind that introducing too complicated clauses may have unforeseen shortcomings, as happens in this case below. We add this property (not the predicate) in the definition by conjuncting it to $f_{gen}(s, e)$ to construct the toy expression:

$$f(s, e) = f_{gen}(s, e) \star_L \big(((b \star_L s) \Rightarrow_L e) \Rightarrow 0\big)$$

In the following we show what stakeholder feedback would look like, and how it would let us evaluate this definition based on context-specific beliefs.

A. Example beliefs

As a first step, we provide example AI system outcomes to stakeholders so that they encounter a wide breadth of situations across the real-world context in which systems are expected to operate. For each set of outcomes, we ask them to evaluate discrimination in the range $[0, 1]$. This corresponds to obtaining the *truth value* of discrimination in reference cases. Recall that these are not the truth values of bias, and therefore stakeholders need not worry about reasoning on whether their observed discrimination is acceptable. A running example on measuring race discrimination for a hiring process is presented in Figure 19; this is artificially constructed for demonstration purposes and does not replicate any real-world data or system.

B. Measures

We assume that an axiomatic definition of discrimination is not possible in this setting, for example because stakeholders cannot agree on which predictive properties should be compared between groups. We instead determine which quantitative data properties loosely replicate similar principles as the ones that influence beliefs, for example through workshops involving domain bias experts and stakeholders. In the running example, discrimination is considered to pertain to some (yet partially unknown) combination of disproportional and unequal representation in hiring results. These correspond to the measures $1 - prule$ [168] and Calders-Verwer disparity cv [169].

C. Replicate beliefs

We now create a mechanism that replicates the truth values of discrimination. In the simplest case, this could be an existing measure found in the literature, or a machine learning model trained on the gathered truth values of discrimination. However, more complicated mechanisms that combine the quantities pinpointed in the previous step could be created. In the running example, the measures are combined through a two-layer perceptron neural network with 64 hidden layers, and trained with gradient descent with 10^{-3} learning rate on the five examples until numerical tolerance 10^{-6}. There will likely be small deviations from the reference truth values of discrimination on validation data (for simplicity, in the running example training data are also used for validation). It is important to again involve stakeholders and policymakers in determining which deviations are acceptable, as shown in the last column of the running example. We could, this or the previous step if some deviations were not acceptable, for example, to retrain the two-layer neural network to greater precision, or to rework which quantities it should learn from. The end-result of the above process is a measure of discrimination (or a model approximating such a measure) that mirrors the belief systems of stakeholders.

D. Determine remaining choices

We finally need to decide on what the group membership predicate S entails and its corresponding

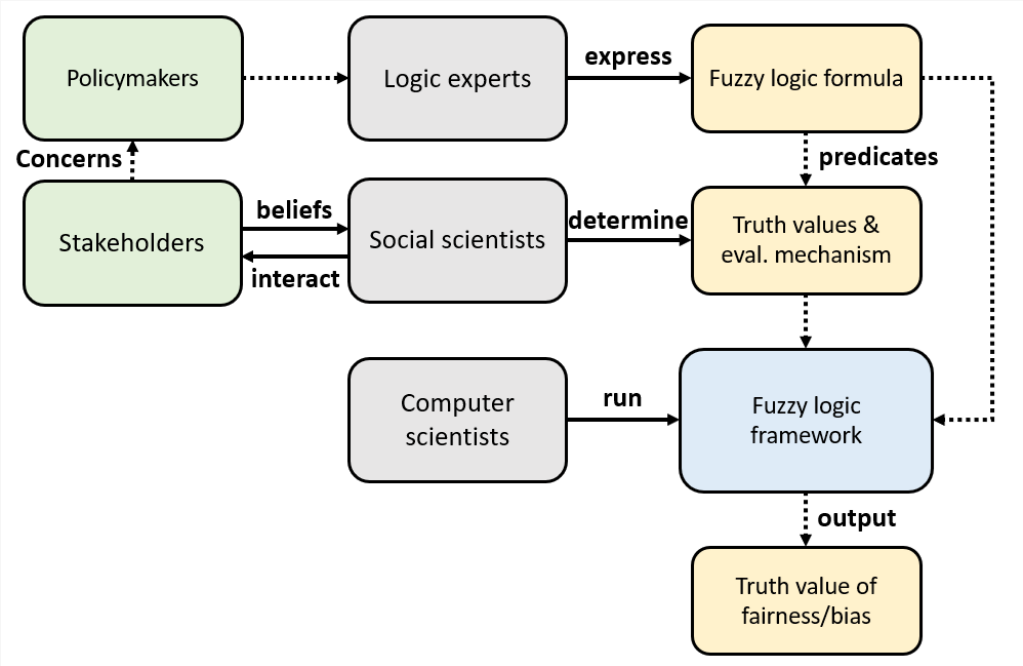


Figure 19. Synethetic example of fitting survey-derived discrimination for truth value of discrimination $P(discrimination) = e$.

truth value, as well as in which subclass of basic fuzzy logic fairness evaluation will take place. In the specific definition of fairness we tackle in this appendix, we also need to determine the truth value of the business necessity predicate B. For the first property, we consider a scenario where interaction with stakeholders led to consider the truth value of encountering one group members as the statistical chance of encountering a person from the group in the real world. This value could be the result of negotiation, but could also be obtained through an additional questionnaire. Let us assume that some hiring policy also set the business necessity predicate to be equal to 10%. To determine which fuzzy logics we should work with, let us assume that, after further negotiation, preserving some minimum kernel of truth value even if multiple statements not perfectly true were encountered was most important for stakeholders. This means that the fairness definition is evaluated in Gödel logic. For this, $f_{gen}(s, e) = 1$.

Numerical example

We finally present an example evaluation of our toy fairness definition: for a set of predictions with $prule = 77\%$, $cv = 0\%$, business necessity 30% and which covers $s = 5\%$ of the population. For these values, our neural network predicts discrimination $e = 19\%$, and therefore

$$\begin{aligned} fairness &= (19\% \star_G 5\% \star_G 1 \star_G ((5\% \star_G 10\% \Rightarrow 19\%) \Rightarrow 0)) \Rightarrow 0 \\ &= (5\% \star_G ((5\% \Rightarrow 19\%) \Rightarrow 0)) \Rightarrow 0 \\ &= 1 \end{aligned}$$

After this first evaluation, we get a sense that the business necessity predicate may be highly contentious in Gödel logic; any value $b > 0$ yields $fairness = 1$ for $s > 0, e > 0$. To avoid similar pitfalls, we stress that properties of fairness definitions written in basic fuzzy logic should

be critically examined. Furthermore, notice how uninformative working with strong negation may be in Gödel (or Product) logic; although it makes it possible to procure perfectly certain fairness, in practice, it does so by discounting a wide range of computational nuance. Thus, it may be preferable to create continuous evaluation by working with bias or its weak negation and only assess fairness at the very end of algorithmic bias mitigation pipelines.

3.6.4 Conclusion

With the goal of quantifying bias and fairness while remaining socially responsible, we investigated how to enable practical adoption of a fuzzy logic framework that can account for policymaker and stakeholder opinions. Our analysis started from an interdisciplinary workflow that makes it possible to actually follow the framework by delegating certain tasks to respective experts and organizing their collaboration. We then explored , whose practical viability we demonstrated with a case study on a toy policy and settings.

3.6.5 Relevant Preprints

- [157] Krasanakis, E., and Papadopoulos, S. (2024). Evaluating AI Group Fairness: a Fuzzy Logic Perspective. arXiv preprint arXiv:2406.18939. *Under review.*

4 T3.3 Cognition-aware hybrid decision-making systems

4.1 Human Cognitive Biases and AI systems

Contributing partners: ALC, FBK

Cognitive biases are systemic patterns of deviation from rationality in decisions made by humans that have been well documented by scholars in social psychology, cognitive science, and behavioral economics since the 1970's [170, 171, 172, 170, 173].

While there is no unified theory of cognitive biases, it is clear that cognitive biases and heuristics – which often lead to sub-optimal outcomes – are a crucial part of our decision making [8]. In fact, they have been commercially leveraged in multiple sectors such as casinos [174], addictive apps [175] and marketing [176, 173] to manipulate human behavior. We advocate for a constructive and positive use of cognitive biases in technology, moving from manipulation to collaboration. We propose that considering our cognitive biases in AI systems could lead to more efficient human-AI collaboration.

There has however been limited research on the interaction between between human biases and AI systems, as recently highlighted by several authors [177, 178, 179, 180]. As a first step, we propose a taxonomy of human cognitive biases that is tailored to the design of AI systems. We then identify 20 cognitive biases that are the most relevant for AI systems. We follow this up with a study on one popular bias – the attractiveness halo effect – and it's impact in the digital age.

4.1.1 A Taxonomy of Cognitive Biases

There exist multiple taxonomies of cognitive biases across specific to particular domains such as medical decision making [181, 182], tourism [183], fire evacuation [184] or visualization[185]. Others classify biases based on their underlying phenomena [186, 187, 188], which could be misleading given that there is no widely accepted theory of the source of cognitive biases [189].

We propose classifying biases according to five stages in the human decision making cycle as depicted in Fig.20 - *presentation* biases, associated with how information or facts are presented to humans; *interpretation* biases that arise due to misinterpretations of information; *value attribution* biases that emerge when humans assign values to objects or ideas that are not rational or based on an underlying factual reality; *recall* biases associated with how we recall facts from our memory and *decision* biases that have been documented in the context of human-decision making. Four exemplary biases in each of these categories along with their relavence for AI systems have been summarized in Tables 12, 13 and 14. Fig.20 also illustrates the three different ways in which AI systems may interact with humans - as external entities in the real world that humans interact with, as active participants working with humans to support decisions or simply as passive observers of human decisions.

Additionally, Table 15 illustrates how AI could potentially provide support in detecting and mitigating some of these biases using the confirmation bias as an example.

4.1.2 Cognitive Biases and AI: Research Directions

Given the ubiquity of AI-based systems in our daily lives (from recommender systems to personal assistants and chatbots) and the pervasiveness of our cognitive biases, there is an opportunity to

[8] While heuristics typically refer to a simplifying rule used to make a decision and a cognitive bias refers to a consistent pattern of deviation in behavior, both terms are used interchangeably in our work as they impact human decisions in a similar way.

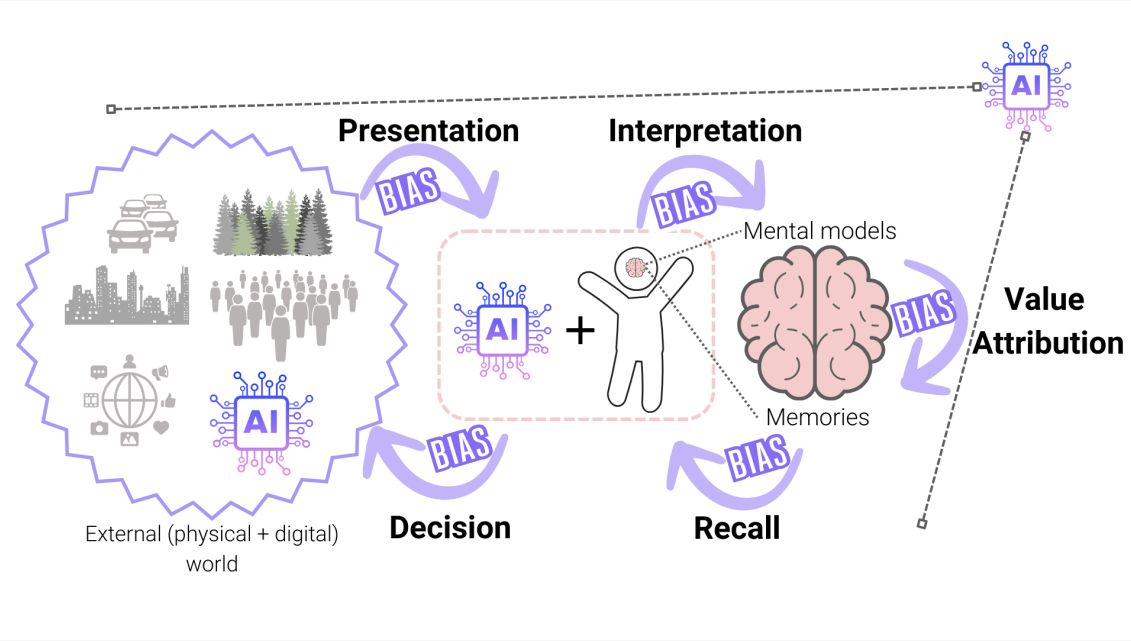


Figure 20. Stages of the human perception, interpretation and decision-making process that are impacted by cognitive biases. AI systems (represented by an orange undirected graph) could observe our behavior, detect biases and help us mitigate them.

leverage cognitive biases to build more efficient AI systems. In our work, we propose three potential directions of research:

1. **Human-AI Interaction:** Older studies of cognitive biases were primarily conducted in scenarios where humans interacted with other humans. However, we do not necessarily perceive interactions with machines in the same way as we do with humans [257]. Thus, it is important to study if cognitive biases are present in interactions with machines, under which conditions and with which degree of intensity. *Presentation biases* from the taxonomy above are the most interesting to study in this setting.

   Beyond verifying their existence in human-AI interaction, these biases could also be used to design more human-like AI systems, which could potentially make them more trustworthy.

2. **Cognitive Biases in AI Algorithms:** Given that humans benefit greatly from relying on decision-making heuristics in certain conditions, it is worth exploring when and how they should be incorporated into AI systems. This could potentially help AI systems learn more from smaller, biased datasets. Initial work by Taniguchi et al. [258, 259] has shown that systems which incorporate the symmetry [260] and mutual exclusion [261] biases can perform better than state-of-the-art methods when the dataset is small and biased. Further research is needed in this area on the potential benefits of other cognitive biases and heuristics.

3. **Computational Modeling of Cognitive Biases:** There is currently work on modeling biases either in a specific domain, such as decisions made by doctors [262, 263], or for a specific task [264, 265]. Bayesian modeling has emerged as a promising modeling framework [178, 266, 179, 267, 268, 269], but it currently cannot model every bias.

A unifying, task-independent AI-based framework to automatically identify cognitive biases from observed human behavior could have a profound impact in the design of AI systems. It could provide insights into unique mitigation strategies and could enable the development of personalized systems for different individuals.

4.1.3 Relevant Publications

- [270] A. Gulati, M. A. Lozano, B. Lepri, and N. Oliver, "BIASeD: Bringing Irrationality into Automated System Design", 2023. https://arxiv.org/abs/2210.01122

4.2 Halo Effect in AI-Driven Beauty Filters

Contributing partners: ALC, FBK

4.2.1 Introduction and Methodology

Building on the work described in Section 4.1, we study the attractiveness halo effect i.e., the tendecy of people to assign positive attributes such as intelligence to more attractive people, even when attractiveness is not a valid cue to predict the dependent variable [220, 223, 271]. We focus on perceptions of intelligence and trustworthiness, given their relevance in human-AI collaboration and human-to-human interaction.

In this task, we use AI-based beauty filters – a popular family of face filters, easily accessible on social media – which aim to *beautify* the face of the person by automatically applying changes to the skin, the eyes and eyelashes, the nose, the chin, the cheekbones, and the lips. They rely on computer vision and augmented reality methods and their prevalence, with millions of users worldwide, profoundly impacts user self-presentation, raising questions about authenticity, self-esteem [272], mental health [273], diversity [274] and racism [275].

4.2.2 Methodology

To study the impact of beauty filters on the attractiveness halo effect, we selected a gender-balanced sample of 462 images from the Chicago Faces Database (CFD) [276] and the FACES dataset [277] that were diverse regarding the age and ethnicity of the faces, while being gender balanced. We refer to this dataset as the `PRI` set. We applied a state-of-the-art beauty filter to each of the images in the `PRI` set to create a new set of *beautified* images (`POST` set) depicting the *same* individuals as in the `PRI` set but in an "attractive" (beautified) condition. We recruited 2,748 participants from Prolific to rate the images on a 7-point Likert scale on perceived attractiveness and 6 other attributes, including intelligence and trustworthiness, as summarized in Fig.21. Each participant evaluated a gender-balanced sample of 10 distinct individuals with an equal number of images from the `PRI` and `POST` sets, but were not told about the beauty filters being applied to some of the images they saw. This large dataset enables us to study the attractiveness halo effect on the same individual in two conditions (original and beautified) minimizing the possibility of factors other than attractiveness impacting the effect; at scale and with diversity in the age, gender and ethnicity of the stimuli, which significantly expands the studies carried out to date [278, 279, 280, 281, 282, 283]. Moreover, we shed light on the impact that pervasive beauty filters have on this cognitive bias.

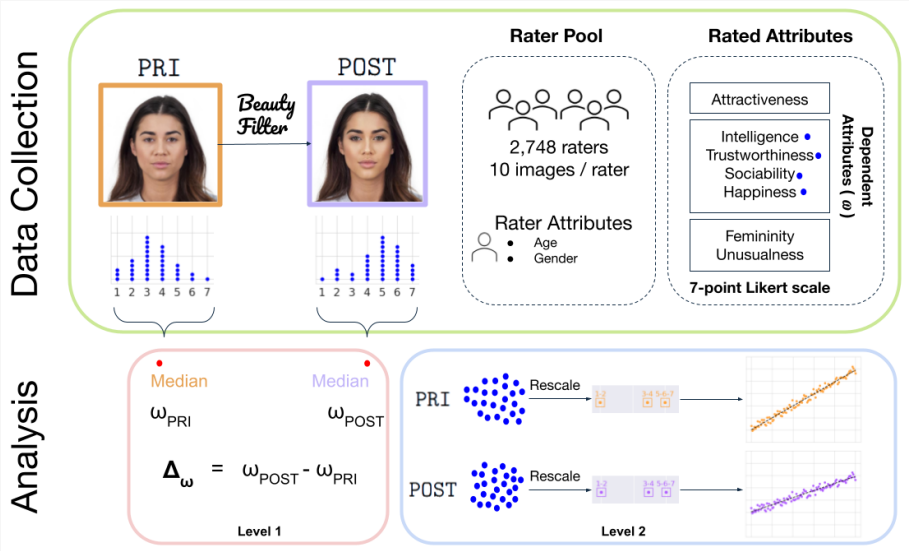
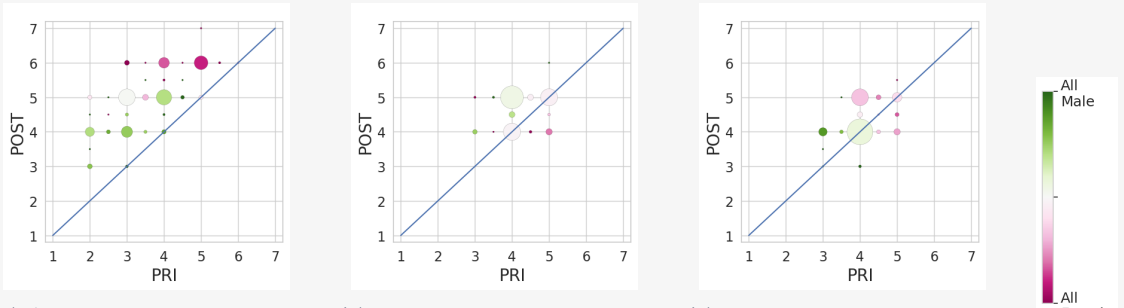


Figure 21. Visual depiction of the user study conducted to collect the data.

4.2.3 Main Findings

Beauty filters increase the perceptions of attractiveness. Comparing the attractiveness ratings of the images before and after the filters were applied revealed statistically significant differences in perceived attractiveness for the same individual before and after beautification ($p < 0.001$, Wilcoxon one-sided pairwise rank test, Fig.22a).



(a) Attractiveness *(b) Intelligence* *(c) Trustworthiness*

Figure 22. Pairwise comparison of (a) attractiveness (b) intelligence and (c) trustworthiness before (x-axis, PRI) and after (y-axis, POST) beautification. The size of the circles is proportional to the number of ratings provided for each value on a 7-point Likert scale and the color indicates the proportion of males and females for each rating. Observe in (a) how all images were rated equally or more attractive after beautification and images of females were rated as more attractive than images of males. Regarding intelligence (b), the images that were rated as less intelligent after beautification were mainly from females.

Gender and age matter. Ethnicity does not. Both before and after beautification, images of females and images of young individuals received significantly higher ratings of attractiveness than images of males or other age groups, respectively ($p < 0.001$, Kruskal-Wallis). Perceptions of attractiveness were not significantly different for subjects across ethnic groups before and after the application of the filters.

Beauty filters mitigate the attractiveness halo effect. While Figures 22b and 22c show that some individuals received lower ratings of intelligence and trustworthiness after beautification, a statistically significant increase in perceptions of intelligence and trustworthiness for the *same individual* was found after the filter was applied, supporting the existence of the attractiveness halo effect ($p < 0.001$, Wilcoxon one-sided pairwise rank tests). However, the halo effect significantly weakened after beautification, as reflected by the parameters of the linear models depicted in Table 16. This result suggests that beauty filters could be used to mitigate the impact of the attractiveness halo effect.

Beauty filters exacerbate gender stereotypes. To study the interaction of the gender and age of the raters and the subjects in the images, we built linear models of intelligence and trustworthiness with attractiveness, the age and gender of the stimuli and the rater, and interactions between these variables as regressors, considering the raters as random effects. We compared the estimated marginal means depending on the gender of the rater and the subjects in the images and identified several gender differences, reflected in Fig.23:

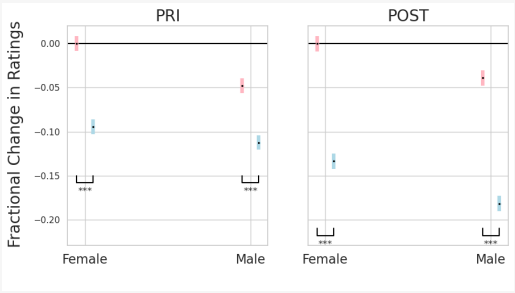
- Images of females received higher scores than males, with the gap expanding after beautification. Male raters also gave lower scores of attractiveness than female raters.
- Despite receiving higher attractiveness scores, images of females received lower intelligence scores than images of males, with the gap again increasing after beautification. Thus, gender played a stronger role in perceptions of intelligence and beauty filters exacerbated this stereotype.
- Regarding trustworthiness, the images of females in the `PRI` set were considered to be more trustworthy by both male ($p < 0.001$) and female raters ($p < 0.001$), yet male raters considered images of males and females to have similar levels of trustworthiness after beautification. Thus, perceptions of male raters seem to be more strongly impacted by the beauty filters.

4.2.4 Conclusion

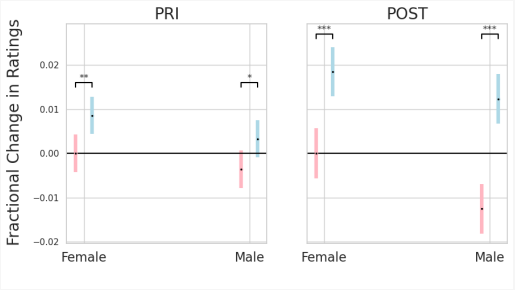
Our findings raise several ethical considerations. First, the fact that females are rated as less intelligent yet more attractive than males corroborates the existence of harmful gender stereotypes and the notion that physical appearance is inversely related to intelligence. Second, we find that male raters are more influenced by beauty filters than female raters, raising concerns about how beauty filters may reinforce existing societal biases, potentially perpetuating gender discrimination and reinforcing traditional gender roles. Finally, while our study suggests that beauty filters could mitigate the intensity of the attractiveness halo effect, their use raises questions about authenticity and honesty. Thus, there is a need for transparency and ethical guidelines surrounding the use of beauty filters, especially in contexts where individuals may be influenced in their decision-making by filtered images without their knowledge. In sum, our findings underscore the complex interplay between technology, social perceptions and ethics, highlighting the need for a deeper understanding of the implications of beauty filters.

4.2.5 Relevant Publications

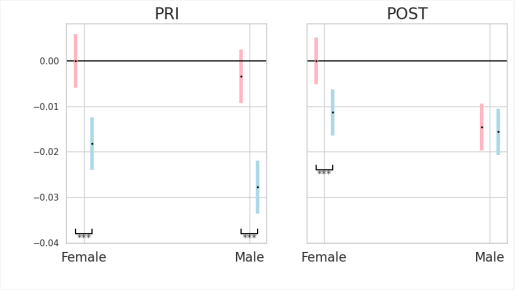
- Gulati, A., Martinez-Garcia, M., Fernández, D., Lozano, MA., Lepri, B., & Oliver, N. (2024). What is beautiful is still good, especially if you are man: The Attractiveness Halo Effect in the era of AI-based Beauty Filters. *International Conference on Computational Social Science.*



(a) Attractiveness



(b) Intelligence



(c) Trustworthiness

Figure 23. Impact of rater's and stimulus' gender on attractiveness and the dependent variables in the `PRI` and `POST` sets. The x-axis represents the gender of the rater and the colours represent the gender of the stimulus (pink [•] for images of females and blue [•] for images of males). The width of the bars corresponds to the 95% confidence interval of the Estimated Marginal Mean (EMM) [7, 8]. The y-axis depicts the relative change in the EMM from the EMM of female stimuli rated by female participants.

- Gulati, A., Martinez-Garcia, M., Fernández, D., Lozano, MA., Lepri, B., & Oliver, N. (2024). What is beautiful is still good: The Attractiveness Halo Effect in the era of AI-based Beauty Filters. *International Conference on Thinking.*

4.2.6 Relevant Software and/or External Resources

- https://github.com/gulu42/theBeautySurveyAnalysis

4.3 Detection of Cognitive Biases in Global News

Contributing partners: JSI

4.3.1 Introduction and Methodology

In this section, we describe the path towards identifying biases related to technologies in global news. Based on experiments conducted by psychologist Peter Wason in 1960s (Wason's rule discovery task), it was demonstrated that people tend to seek information that confirms their existing beliefs [284]. In our research work we intend to explore the possible confirmation bias related to technologies, along with several other biases:

- Negativity bias: the tendency for the brain to subconsciously place more significance on negative events than positive ones.

- Proximity bias: is the subconscious tendency to give preferential treatment to people/events that are physically close. As an example, a person employee being considered for a raise before a remote employee because they are in the immediate vicinity of their superior is an example of proximity bias.
- Recency bias: the tendency for the brain to subconsciously place more value on the last information it received about a topic.
- Confirmation bias: The tendency for the brain to value new information that supports existing ideas.

Specifically, the news analysis is based on multi-lingual and cross-lingual data obtained from the Event Registry (ER) service [285]. ER is the world's leading news intelligence platform, empowering organizations to keep track of world events and analyze their impact. ER provides content from over 150,000 news sources in 40+ languages, minutes after it has been published. Fig.24 demonstrates the user interfacte of ER service. One of the benefits of using ER service for detection of bias in news is easy to integrate news API to obtain access to real-time as well as archive news content.

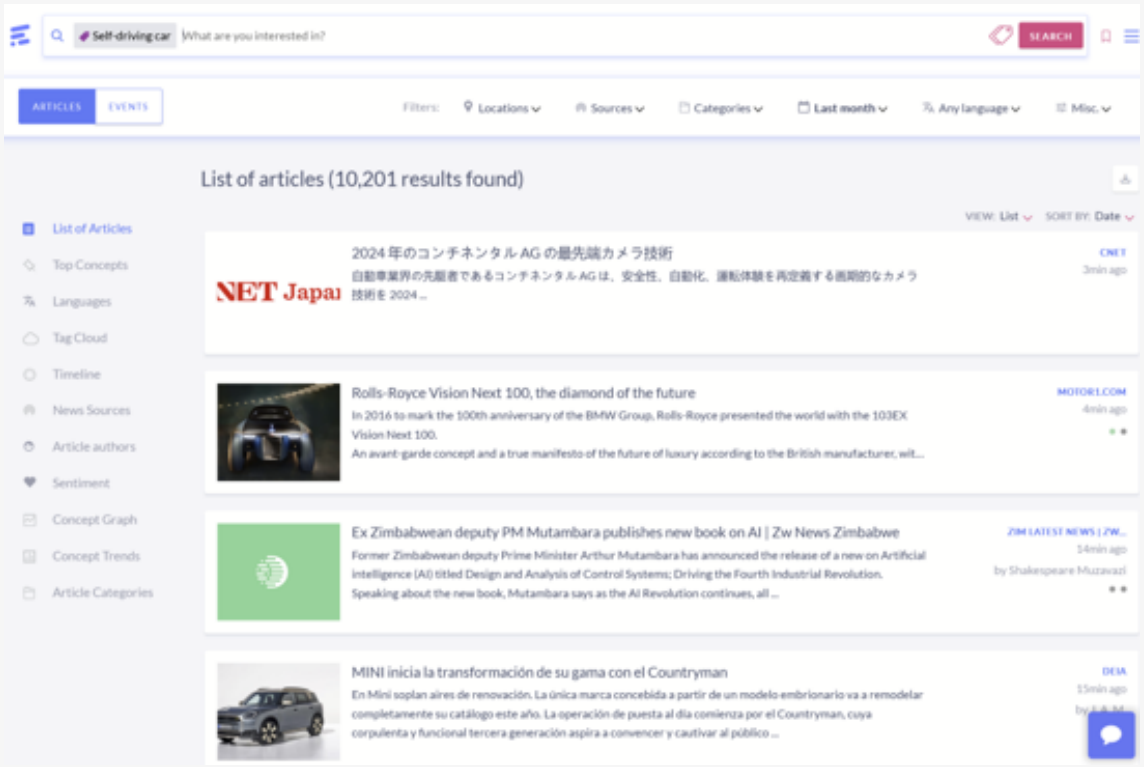


Figure 24. Interface for Event Registry service

The methodology for bias detection in news is based on several steps, such as:

- Obtaining set of news related to Artificial Intelligence technologies (specifically related to incidents associated with AI technologies). For this purpose we have used semantic annotation techniques and the InnoGraph Artificial Intelligence Taxonomy [286].Artificial Intelligence taxonomy uses Wikipedia articles (and categories) as the core part, since every important

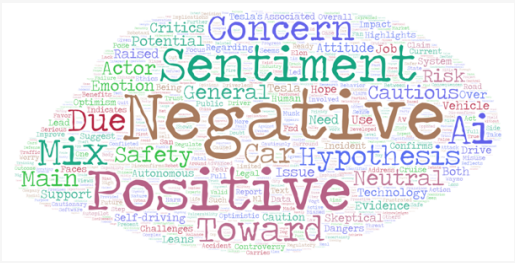
Figure 25. AI technologies: ontology concept example.

topic related to AI (old, new, emerging, popular, etc.) will most likely have an article on Wikipedia that can serve as a starting point. Wikipedia categories are used to classify articles, and to form a hierarchy. They are the result of a collective curation effort to create a taxonomy of Wikipedia articles. Categories are used as a way of finding relevant articles. The current taxonomy version contains around 7.000 concepts and is regularly maintained.

- We have used filtering for specific topics (based on the taxonomy concept). Fig.25 provides an example of ontology concept used for news filtering.
- In current wave of experiments LLM techniques have been applied for identification of specific information in news.
- The produced initial results have been analyzed and prepared for future exploration.

4.3.2 Experiments and Results

In the current experimental setting we have analyzed 2,753 news items related to ”self-driving cars” and with LLMs (specifically, ChatGPT 3.5) identified the following information in each news:



(a) Initial sentiment analysis

(b) Initial emotions analysis.

- actors (name and type)
- technologies (name and type)
- relations (actor, technology, relation type)
- hypothesis (main hypothesis, confirming evidence, disconfirming evidence, sentiment, emotion and actor attitude toward hypothesis)

The example of main hypothesis: "Apple hopes to build a truly autonomous vehicle intelligence". Confirming evidence: "Apple's approach to deploy intelligence at the edge rather than in the cloud". Disconfirming evidence: "Safety drivers taking control of experimental self-driving cars approximately one time for every mile driven". Following that, we have analyzed the sentiment and emotions related to main hypothesis from news. The results are visualized on Figs. 26a and 26b. The further steps would include exploration of the outcomes related to agents as news sources, detection of signals over time and exploration of proximity signals for technologies from news.

4.3.3 Conclusion

In conclusion, the current experiments confirm the applicability of LLMs for bias identification in news. However, more experiments are required to identify and build the appropriate and useful strategies for different types of biases.

4.3.4 Relevant Publications

- I. Novalija, L. Rei, J. Pita Costa and M. Grobelnik. "Identifying Bias Related to Technologies in News", SIKDD Conference Proceedings 2024, October 2024 (in preparation)
- J. Pita Costa and I. Novalija. "Exploring data bias with large language models, moving the sustainable development goals forward", proceedings of the Anna Lindh Foundation policy dialogue in the Euro-Mediterranean region with their evidence-based research or relevant practical experience, 2024 (in preparation)

- M. Jermol and J. Pita Costa, ”Onwards to an Ethical Education for Sustainability through AI”, special issue of the Journal of Artificial Intelligence and Sustainable Development (in preparation)

	Bias	Brief Description	Relevance to AI
Presentation	*Decoy effect* [190, 191, 192, 193]	Placing deliberately a worse alternative between two choices can reverse the user's preference	Could AI systems learn to place decoys effectively while presenting alternatives? Could AI systems learn to identify decoys? [194]
	Framing effect [195, 196, 197, 198]	How a statement is framed can alter its perceived value	Studies have shown that when humans are placed in human-AI teams, their decisions [199] and trust [200] are impacted by the framing effect. Could AI systems learn to frame explanations to make them more trustworthy?
	Anchoring effect [186, 201, 202]	Human decision making is influenced by certain reference points or anchors	The use of anchors to alter user preferences has been studied in marketing and recommender systems [203]. Could AI systems automatically identify anchors that humans might be subject to?
	Pseudocertainty effect [195, 204, 205]	Humans incorrectly estimate the certainty of statements in a multi-stage decision making process	Could AI systems identify situations where humans are likely to be unable to accurately compute the "complete picture"? Could this effect be leveraged by AI algorithms to learn effectively from smaller datasets?
Interpretation	*Conjunction* fallacy [206, 207, 208, 209]	In certain situations, humans see the conjunction of two events as being more likely than any one event individually	Could AI systems recognize situations where humans are likely to make such errors and provide alternate decisions?
	Base Rate fallacy [210, 211]	Humans have a tendency to ignore the base rate information when making decisions	Human reasoning does not follow Bayesian reasoning in certain settings. Could AI systems leverage these non-Bayesian computations effectively?
	Gamblers fallacy [186, 212, 213, 214]	Humans tend to overvalue the impact of past events when predicting the outcome of independent future events	Decision making systems that learn from human decision making –e.g. legal, college admissions or HR decision-making systems– learn from data that reflects the gamblers fallacy. How could this bias be mitigated to design fairer AI-based decision-support systems?
	Hyperbolic discounting effect [215, 216, 217]	Humans tend to choose immediate rewards over rewards that come later in the future	Studies have shown a link between high social media usage and hyperbolic discounting leading to unhealthy behavior [218, 219]. Could AI systems recognize when we are impacted by this bias and help mitigate it?

Table 12. Selected biases in Presentation and Interpretation categories of the proposed taxonomy and their relevance to the study of AI systems.

	Bias	Brief Description	Relevance to AI
Value Attribution	*Halo* effect [220, 221, 222, 223]	Positive attributes associated with a person in one setting carry over other settings	Could this effect be utilized to create systems that are easier to trust? Does the halo effect manifest itself when humans interact with chatbots or robots?
	IKEA effect [224, 225, 226]	Humans associate a higher value to their own creations than those of others	Could this effect be leveraged to provide explanations that users are more likely to accept?
	Risk aversion bias [227, 228, 229, 230]	We tend to avoid risky decisions even if they have a higher net expected utility than less risky options, especially if the uncertainty is high	Could AI systems support human decision-making by counter-balancing the risk aversion risk?
	Social desirability bias [231, 232, 233, 234]	Humans tend to provide the answers to surveys or questions that they believe are expected from them	Do people provide socially desirable answers even when they are interacting with or being evaluated by machines? If yes, could the social desirability bias be be leveraged to nudge users to improve their behavior?
Recall	*False memory* bias [235, 236, 237]	Humans incorrectly remember a past event depending on the questions they are asked about the event	False memories impact how we make decisions. Positive false memories have been shown to have positive consequences [238]. Could this AI systems use this effect to improve user experience?
	Self-reference effect [239, 240]	Events with a direct impact are more likely to be remembered	Could AI systems leverage this effect to make explanations about their behavior more "memorable"?
	Serial-positioning effect [241, 242, 243]	Items at the start and the end of a list are more memorable than those in the middle	When providing explanations, could AI systems leverage this bias to have a more effective human interaction?
	Peak-end rule [244, 245, 246]	The value of an event tends to be assessed based on its peak and final values, neglecting other parameters, such as its duration or average value	Could this bias be leveraged by AI systems to increase the perceived utility of hard tasks? Instead of maximizing the time a user stays on a social media platform, could this bias be used to reduce time spent while also increasing user satisfaction?

Table 13. Selected biases in the Value Attribution and Recall categories of the proposed taxonomy and their relevance to the study of AI systems.

	Bias	Brief Description	Relevance to AI
Decision	*Status quo* bias [247, 248]	Humans tend to make decisions that maintain the current state rather than changing it	Are AI systems that recommend fewer changes more likely to be trusted?
	Shared information bias [249, 250, 251]	In group settings, humans tend to focus the discussion on information everyone already has rather than trying to bring in new information	Could AI systems leverage this bias to effectively drive group conversations?
	Naive allocation [252, 253, 254]	People tend to allocate resources equally between all options rather than based on the value of the options	Could AI systems provide alternatives to avoid naive allocation? Alternatively, could AI systems leverage this bias effectively to make decisions in high uncertainty situations?
	Take-the-best heuristic [255]	The decision between two alternatives is made based on the first cue that discriminates them	Could this heuristic be implemented in AI systems to learn efficiently from small datasets? [256]

Table 14. Selected biases in the Decision category of the proposed taxonomy and their relevance to the study of AI systems.

Cognitive Bias	`Confirmation Bias`
Definition	The tendency of individuals to seek, interpret, and remember information in a way that confirms their preexisting beliefs or hypotheses, while ignoring or dismissing contradictory evidence.
Example	Sam believes that a certain dietary supplement has remarkable health benefits. They regularly read online forums and articles that promote the supplement's positive effects, and they tend to ignore or dismiss any information suggesting potential risks or ineffectiveness.
AI's Role	*Detecting the bias*: Through the collection and analysis of both relevant online content and Sam's information consumption behavior related to the dietary supplement, machine learning algorithms could be used to detect the existence of the confirmation bias. *Counteracting the bias*: Machine learning-based recommendation algorithms could recommend diverse perspectives and evidence-based information to challenge Sam's preconceived notions and encourage a more balanced understanding of the supplement's properties.

Table 15. Example of how AI could support humans in mitigating the confirmation bias.

Dependent Attribute (ω)	PRI			POST		
	β_0	β_1	R^2	β_0	β_1	R^2
Intelligence	3.18***	0.30***	0.327	4.11***	0.12***	0.036
Trustworthiness	3.34***	0.20***	0.181	3.50***	0.17***	0.069
Sociability	2.56***	0.39***	0.363	2.78***	0.38***	0.321
Happiness	2.08***	0.39***	0.261	2.47***	0.35***	0.186

Table 16. Parameters of the linear model $\omega = \beta_0 + \beta_1 Attrac + \epsilon$ for each dependent variable ω on the PRI and POST sets independently. A larger absolute value of the intercept β_0 in the POST set indicates that the value of the perceived attribute increases after applying a beauty filter. A smaller absolute value of β_1 in the POST set reflects a weaker halo effect after beautification.

5 T3.4 Privacy-Preserving Machine Learning

5.1 Model-Agnostic Federated Learning with a Privacy Perspective

Contributing partners: ALC

5.1.1 Introduction and Methodology

Federated Learning (FL) has been proposed as a privacy-preserving solution for machine learning [287]. In FL, the sensitive data never leaves the device of the holder. The idea is that instead, the holders, called clients train their own machine learning models. In the scenarios it is used, the clients do not have enough data to train a good model alone. Thus, the clients collaborate with each other by sharing the model parameters with a server that aggregates these parameter updates to generate a global model. Then, the server can broadcast the global model that benefited from the knowledge of each client. These steps together are called a training round and a FL training can consist several round until it converges to a model that solves the given task.

Recent works have reported that FL can leak private client data through membership inference attacks conducted on the parameters shared between clients and server to retrive information about the original training data [288, 289]. For traditional machine learning it has been shown that the effectiveness of these attacks on the model negatively correlates with the complexity of the model [290].

With this motivation in mind, the work in this task focuses on analyzing the privacy and performance implications of FL methods that leverage the use of different model complexity on the clients. We aim to use smaller model complexity on clients that have higher privacy risk. We investigate the effect of chosing algorithms to project the server's larger model to the clients' smaller ones.

Model-agnostic federated learning The simplest form of federated learning is the method called federated average (FedAvg) [287]. In `FedAvg`, the clients train the same model as the server using their own dataset, such that the average of the clients' model weights is an approximation of training the same model in a centralized machine with access to all client data. That is, `FedAvg` computes $\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$, given by:

$$\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \frac{1}{|\mathbb{D}|} \sum_{c=1}^{C} \sum_{(\boldsymbol{x},y)\in\mathbb{D}_c} l(y, f(\boldsymbol{x}, \boldsymbol{\theta})) \approx \frac{1}{C} \sum_{c=1}^{C} \min_{\boldsymbol{\theta}_c} L_c(\boldsymbol{\theta}_c, \mathbb{D}_c) \tag{16}$$

where L is the loss function in the server when having access to all the client data; l is the loss function in each client; $\boldsymbol{\theta}$ and f are server model parameters and server architecture, respectively. The loss at each client $L_c(\boldsymbol{\theta}_c, \mathbb{D}_c)$ is given by $\frac{1}{|\mathbb{D}_c|} \sum_{(\boldsymbol{x},y)\in\mathbb{D}_c} l(y, f(\boldsymbol{x}, \boldsymbol{\theta}_c))$, where C is the number of clients; $\boldsymbol{\theta}_c$ is the client model parameters; and $\mathbb{D}_c$ represents the dataset of client c such that $\mathbb{D} = \bigcup_{c=1}^{C} \mathbb{D}_c$ corresponds to the entire dataset. In this example, the model architecture f is shared between clients and server.

A Federated Learning architecture is called *model-agnostic* when the machine learning model architecture is different –typically smaller– in the clients than in the server. To enable the federation, one can leverage knowledge distillation [291, 292], or perform parameter aggregation if the clients share a subset of the server's model architecture $f_c \subset f$ [293, 294]. For example, if the federation performs a computer vision task using CNNs both in the server and the clients, this translates to a smaller number of input and output channels in clients with smaller model complexity than the server's. Our research investigates this latter scenario.

Membership inference attacks (MIAs) While FL was initially motivated by the desire to preserve client data, recent studies have revealed that federated systems remain vulnerable to privacy attacks, specifically in the form of membership attacks [295, 296, 297, 298]. In our work, we focus on membership inference attacks (MIAs) [299], where the attacker's goal is to determine whether an individual data point was part of the dataset used to train the target model. While MIAs expose less private information than other attacks, such as memorization attacks, they are still of great concern as they constitute a confidentiality violation [300]. Membership inference can also be used as a building block for mounting extraction attacks for existing machine learning as a service systems [289]. Several types of MIAs have been proposed in the literature [301, 302, 303].

Attacks can occur on the client or the server-side: (1) *client* exposure or attack occurs when the attacker targets the client model, $(f_c, \boldsymbol{\theta}_c^t)$, for client $c = 1, \ldots, N$ in training round $t = 1, \ldots, T$. In a stateful setting [304], the attacker can collect a set of $k \leq T$ client updates in $\Theta_c = \{\boldsymbol{\theta}_c^{\tau_1}, \ldots, \boldsymbol{\theta}_c^{\tau_k}\}, \tau_i \in \{1, .., T\}$ training rounds; (2) *server* exposure or attack takes place when the attacker is able to listen to the parameter $\boldsymbol{\theta}^t$ updates that are broadcasted by the server to the clients. The attacker aims to identify the entire training dataset $\mathbb{D} = \sum_c \mathbb{D}_c$. In this paper, we consider *black-box, passive, client-side* attack on the last update sent from the client to the server $\boldsymbol{\theta}_c^T$ where the attacker aims to identify instances of the client's dataset $\mathbb{D}_c$ for client c.

In a black-box attack, the attacker has no direct access to the model's parameters $\boldsymbol{\theta}_g$ and architecture f, but it can query the model with data instances to get the model prediction $\hat{y}$. The attacker's purpose is to build an attacker model $\mathcal{A}$ that predicts, for data instance $(\boldsymbol{x}, y)$, if it was part of the training data $\mathbb{D}_g$ of model $M(f, \boldsymbol{\theta}_g, \mathbb{D}_g)$, where the subscript g can denote both the server and each of the clients. Passive attackers observe the behavior of a system without altering it, while active attackers engage with the system by modifying inputs or parameters to exploit vulnerabilities or extract information.

Formally, the perfect attacker's model $\mathcal{A}$ is given by:

$$\mathcal{A}(f, \boldsymbol{\theta}_g, (\boldsymbol{x}, y)) = \begin{cases} 1, & \text{if } (\boldsymbol{x}, y) \in \mathbb{D}_g, M(f, \boldsymbol{\theta}_g, \mathbb{D}_g) \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

In this work, we study the performance of three different black-box, membership inference attacks:

- `Yeom` **attack:** In this light-weight, loss-based attack [290], the attacker chooses a global threshold ν, and selects every data instance with a loss lower than ν as a member of the training dataset.
- `LiRA` **attack:** In the offline version of the attack of [305], the attacker has an auxiliary dataset $\mathbb{D}_a$ and trains shadow models $M_{sw}(f, \mathbb{D}_{sw})$ on random subsets of this dataset $\mathbb{D}_{sw} \subset \mathbb{D}_a$. The data instance is predicted to be a member of the client's training set if the target model's confidence score fits into the sample's confidence score distribution in the shadow models.
- `tMIA` **attack:** a state-of-the-art attack that uses knowledge distillation to collect loss trajectories to identify member and non-member instances [306]. The method builds on the idea that the snapshots of the loss after each training epoch (loss trajectory) can separate the member instances from non-members better than only using the final model's loss.

5.1.2 A Taxonomy of Channel Selection Algorithms

In this research, we explore the field of model-agnostic federated learning with shared model weights $(f_c \subset f)$ in CNNs. In this context, two popular methods are `FDropout` and `HeteroFL`.

In FDropout [293], all the clients learn a CNN with the same architecture but fewer parameters (smaller weight matrices) than the server, and the server randomly drops a fixed number of units from each client [307], mapping the sparse model to a dense, smaller network by removing the dropped weights.

HeteroFL [294] follows a similar idea as FDropout but with two key differences when selecting the channels in the clients with smaller models than the server: 1) all the clients learn from the same portion of the server's model; and 2) instead of randomly dropping cells, the clients always keep the top-left subset of the server's weight matrix for each layer in the network. Thus, in HeteroFL, the weight matrix $\boldsymbol{A}_c^l$ of size $N_c \times M_c$ in layer l and client c corresponds to the top-left sub-matrix of the server's weight matrix $\boldsymbol{A}^l$ of size $N \times M$.

We propose a taxonomy of channel selection methods that reveals the similarities of these two methods and enables us to identify gaps which we fill by implementing seven more model-agnostic FL algorithms. Figure 27 presents the proposed taxonomy and its three dimensions.

The first dimension of the taxonomy refers to coverage, classifying the methods in three classes: *one group* (O); *several groups* (G); and *unique* (U), depending on the number of channel sets used to train in the clients with smaller models than the server's model. In *one group*, each client selects the same set of channels. In *several groups*, clients are clustered in groups such that clients in the same group use the same set of channels (Figure 27 shows an example with 4 groups). *Unique* corresponds to federations where every client has their set of channels selected individually.

The second dimension characterizes the strategy for channel selection and defines two types: *fixed* (F) methods when the channel sets are defined at the beginning of the training, and *resampled* (S) methods when the channel sets are selected in each training round.

Finally, the third dimension divides methods into two kinds: *submatrix* (M) methods if the selected channels are the first or second half of the full channel list and *random* (R) methods if the channels are selected randomly.

According to the proposed taxonomy, FDropout corresponds to a USR method because each client has a different, random set of channels in each training round, while HeteroFL corresponds to an OFM method as there is only one client group with fixed channels that correspond to a sub-matrix of the server's weight matrix.

Each of the proposed methods is anticipated to provide a different privacy-accuracy trade-off, depending on how the clients select the portions of the server's matrix to use in their training. We formulate below three hypotheses in this regard that we validate empirically in our experiments.

H1: Frequency Hypothesis. We hypothesize that methods where clients have access to the same set of channels more frequently perform better in terms of client model accuracy but have a worse client-level privacy. For example, in GFM the clients access the same set of channels every four rounds. Thus, compared to HeteroFL and FDropout, we expect this method to yield a client privacy-performance trade-off between these two existing methods.

Based on this hypothesis we expect:

- OSR, GSR, and USR (i.e., FDropout) to be the most resilient methods against MIAs but provide the worst client accuracy as the clients receive the parameters from a new set of channels in every round. Therefore, the same set is only repeated in every $\binom{N}{N_c}$ rounds on average for client c with client channel size N_c and server channel size N.
- OFM (i.e., HeteroFL) and OFR to be the most vulnerable against MIAs but achieve high client accuracies as the clients train using the parameters of the same set of channels in every round (1 round).

H2: Similarity between the M and R categories. In a CNN layer, as long as the selected input channels of layer l match the output channels of layer $l-1$, the differences between variations

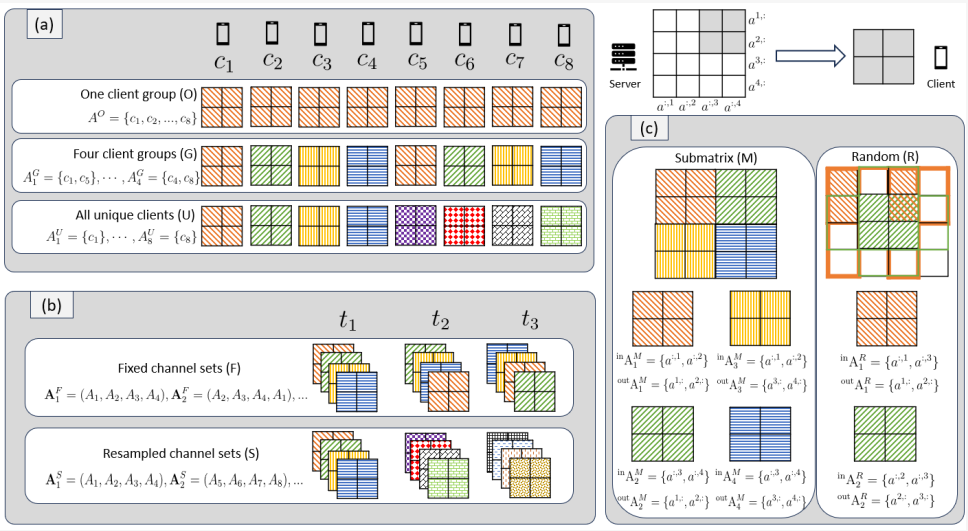


*Figure 27. Taxonomy of channel selection methods for model-agnostic FL architectures with fixed model size in the clients. The server and the clients learn the same type of models (e.g. CNNs) but with different numbers of units. The clients with smaller models need to select channels from the server's model as part of their learning process. The taxonomy considers three dimensions: **a)** Number of clients learning from the same server channels: one group (O), four groups (G), all unique clients (U); **b)** Channel group selection: fixed at the beginning of the training (F), sampled in each round (S); and **c)** Channel selection: according to a submatrix structure (M), randomly (R).*

M and R should be small. They differ only in the number of channels shared by client groups. We designed the sub-matrix category (M) to minimize the channel overlap between groups. Thus, we expect the models in the M and R categories to behave similarly.

H3: The differences in the privacy-accuracy trade-off between the methods decrease as the number of large clients in the federation increases The channel selection methods discussed in this paper are relevant when the majority of the clients learn smaller models than the server's model. In fact, in cases when all the clients but one learn models of the same complexity as the server's model, the `UFR` and `OFR` methods become the same. Therefore, we expect the impact of the channel selection strategies to be larger when the majority of clients in the federation learn smaller models than the server's model.

We perform a comparative analysis of the proposed model-agnostic FL methods and empirically validate our hypotheses in experiments on two vision datasets, as described next.

5.1.3 Experiments and Results

We conducted experiments on vision datasets CIFAR-10, CIFAR-100, and FEMNIST simulating 10 clients in federated learning settings with a simple convolutional neural network (CNN) with 4 convolutional layers and one fully connected layer at the end following [294].

Model complexity, privacy and dataset size. Previous work has shown that as models get more complex, they are more vulnerable to MIAs. For example, [290] demonstrate that their attack's accuracy increases as the model size increases on standard benchmark image datasets.

In Federated Learning, [308] reported that, the larger the models, the more vulnerable they are to model memorization attacks. In their case, it was a horizontal FL architecture with the same model (ResNet) both in the server and the clients. Other works have highlighted that over-

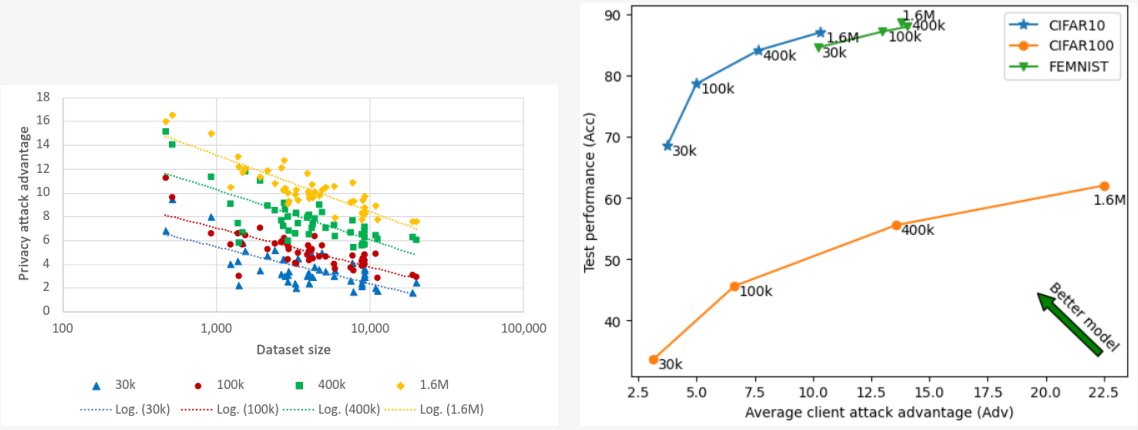
parameterized models are vulnerable to membership memorization attacks [309].

In this experiment, we shed further light on this topic by focusing on the privacy-accuracy trade-off in FL with respect to dataset and model size, and from the perspective of both the server and the clients. Note that prior studies have only analyzed the server's performance.

By means of an empirical illustrative example, we show that, for a given model and an FL scenario, there is a strong negative correlation between the size of the clients' datasets and models, and their vulnerability against membership inference attacks (`Yeom` in our example).

As previously discussed, this attack occurs on the last update the client sends to the server in round T, $\mathcal{A}_{\texttt{Yeom}}(\boldsymbol{\theta}_c^T)$. We use the `Yeom` attack for this illustrative experiment as it requires significantly less computation than the other described MIAs. We report the **attack advantage** [310], the improvement of an attack when compared to the baseline random guess on a balanced test set, that yields 50% attack accuracy. Therefore the attack advantage is $Adv(\mathcal{A}) = 2(Acc(\mathcal{A}) - 50)$, where $Acc(\mathcal{A})$ is the accuracy of the attacker's model.

Figure 28 describes the results of this experiment. We observe strong negative correlations between the size of the clients' dataset and the attack's advantage; and between the clients' model complexity and the corresponding attack's advantage. We also observe that both the attacker's advantage and the test accuracy on the clients increase as the model size increases. These results suggest that model-agnostic FL could enhance privacy both in the server and the clients by means of learning models in the clients that are smaller than the server's model.



(a) *(b)*

Figure 28. ***(a)****: Exemplary illustration of the correlation between the privacy attack advantage for the* `Yeom` *attack and the dataset size from the clients' perspective. Results for 5 repeated experiments on the CIFAR-10 dataset using the* `FedAvg` *architecture with 10 clients having different dataset sizes, resulting in 50 client models. Each dot depicts a client in one federated training and the color represents different model complexities (CNNs), characterized by the number of parameters, ranging from 30k to 1.6 million. Note the negative correlations between the size of the clients' dataset and the attack advantage, as well as between the model's complexity and the associated attack advantage.* ***(b)****: Privacy-accuracy trade-off of the data depicted in (a) by averaging experiments across clients per model complexity. In addition to CIFAR-10, we also show the trade-off for the CIFAR-100 and FEMNIST datasets. The attacker's advantage and test accuracy on the clients increases as the model size increases. Observations in (a) and (b) suggest that model-agnostic FL could be a privacy-enhancing solution.*

Privacy of model-agnostic methods We train the 9 model-agnostic channel selection methods in a FL architecture with 10 clients, of which 2, 5, or 8 clients learn small models and the rest learn models of the same complexity as the server's model. Clients with a smaller dataset size are selected first to learn smaller models. We also train two `FedAvg` baselines where the server and all

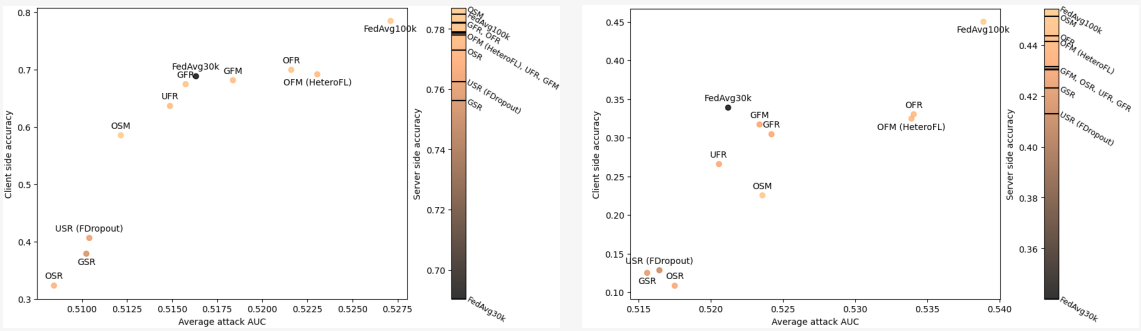
the clients learn models with the large and the reduced model sizes. This results in training 29 FL models for each data distribution.

Each client is subject to the 3 previously described MIAs. For `LiRA` and `tMIA`, the auxiliary dataset is drawn from the data of the rest of the clients $\mathbb{D}_a^c = \{\mathbb{D}_1, ..., \mathbb{D}_C\} \setminus \mathbb{D}_c$. We use the same shadow models to attack models from the same experiment. We train 16 shadow models for `LiRA` and use 25 distill epochs for `tMIA`.

Figure 29 depicts the average attack AUC of the 3 MIAs on all the model-agnostic FL channel selection methods with 2 *large* clients. Experiments on CIFAR-10 and CIFAR-100 corroborate our first hypothesis *H1* related to the accuracy-privacy trade-off. From a client perspective, methods `GFM` and `GFR` achieve similar results to the `FedAvg30k` baseline with small models, yet, they outperform `FedAvg30k` on the server-side accuracy (77.89% and 78.19% over 69.04% for CIFAR-10 and 43.05% and 43.17% over 34.01% for CIFAR-100). Furthermore, their privacy compared to `HeteroFL` is better by $0.5 - 1.0\%$ AUC with very similar levels of client accuracy. `FDropout`, and the `GSR` and `OSR` methods perform well in terms of client privacy, but their client accuracy is significantly lower when compared to the rest of the methods.

Supporting our *H2* hypothesis, methods `GFR` and `GFM`, and methods `OFR` and `OFM` yield similar results in all three measures on the two datasets, with `OFM (HeteroFL)` and `OFR` on CIFAR-100 being the closest with differences of only $0.2\%, 0.6\%$, and 0.02% on the server accuracy, client accuracy, and attack AUC, respectively.

Interestingly, while the `OSM` method performs as expected on the client side, it outperforms every method on its server-side accuracy. Therefore, it provides the best server accuracy-client privacy trade-off from all the studied model-agnostic FL methods.



(a) *(b)*

Figure 29. Experiments to measure the client-side accuracy, server-side accuracy and the client's privacy. MIA vulnerability averaged over the 3 attacks (`Yeom`, `LiRA`, `tMIA`). All model-agnostic FL architectures have 2 clients with large models and 8 clients with small models. FedAvg30k and FedAvg100k are the baselines with 10 small (30k parameters) and 10 large (100k parameters) clients respectively.

5.1.4 Conclusion

Our findings can be summarized in three points. Firstly, **model compression helps against MIAs in FL**. We showed that the relation between the model complexity and privacy that was observed in traditional machine learning holds in FL and FL with agnostic models. Secondly, we constructed **a novel taxonomy of heterogeneous FL methods** and proposed seven new models. Lastly, **the channel selection strategy in heterogeneous FL matters** in terms of client-side privacy, accuracy and server-side accuracy.

5.1.5 Relevant Publications

- [311] G. D. Németh, M. Á. Lozano, N. Quadrianto, and N. Oliver, "Addressing membership inference attack in federated learning with model compression," arXiv preprint arXiv:2311.17750, 2023.

6 Ongoing Work and Conclusions

This deliverable report provides a detailed overview of the progress achieved by M12 in tasks T3.2, T3.3 and T3.4.

Concerning Task T3.2, the report focuses on advancements in fairness and counterfactual interventions, covering a range of topics such as bias mitigation in multimodal systems, fairness measurement using Shapley values, removal of toxic content from vision-language models, early detection of hallucinations in diffusion models, and the application of fuzzy logic to quantify fairness. Task T3.3 explores the presence of cognitive biases in AI systems and human biases triggered by AI-generated beauty filters, as well as methods to detect cognitive biases in news media. Task T3.4 addresses privacy-preserving machine learning, with an emphasis on federated learning, showing that larger datasets and more complex models improve resistance to membership inference attacks. The report also discusses the potential of model-agnostic federated learning to enhance privacy by allowing clients to utilize models with varying complexities.

Overall, the activities carried out in this work package have been substantial, including eight papers published in top conferences and journals, two preprints, and two additional papers currently in preparation. These accomplishments are well distributed among the contributing research groups, reflecting their active engagement and commitment to the objectives of the work package. The significant number of publications and ongoing work attests to the efforts of the partners in advancing the project's goals. This deliverable covers only the initial results; however, the current trajectory indicates that we are progressing well in line with the original plan.

References

[1] Z. Wang, Z. Sha, Z. Ding, Y. Wang, and Z. Tu, "TokenCompose: Grounding Diffusion with Token-level Supervision," *arXiv preprint arXiv:2312.03626*, 2023.

[2] F. Friedrich, M. Brack, L. Struppek, D. Hintersdorf, P. Schramowski, S. Luccioni, and K. Kersting, "Fair diffusion: Instructing text-to-image generation models on fairness," *arXiv preprint*, 2023.

[3] C. Zhang, X. Chen, S. Chai, C. H. Wu, D. Lagun, T. Beeler, and F. De la Torre, "Iti-gen: Inclusive text-to-image generation," in *ICCV*, 2023.

[4] P. J. Kenfack, K. Sabbagh, A. R. Rivera, and A. Khan, "Repfair-gan: Mitigating representation bias in gans using gradient clipping," *arXiv preprint*, 2022.

[5] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving latent diffusion models for high-resolution image synthesis," in *ICLR*, 2024.

[6] P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting, "Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models," in *CVPR*, 2023.

[7] S. R. Searle, F. M. Speed, and G. A. Milliken, "Population marginal means in the linear model: An alternative to least squares means," *The American Statistician*, vol. 34, p. 216–221, Nov. 1980.

[8] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2023. R package version 1.9.0.

[9] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *NeurIPS*, 2022.

[10] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *International Conference on Machine Learning*, 2022.

[11] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint*, 2022.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[13] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *CVPR*, 2023.

[14] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint*, 2022.

[15] D. Epstein, A. Jabri, B. Poole, A. A. Efros, and A. Holynski, "Diffusion self-guidance for controllable image generation," in *NeurIPS*, 2023.

[16] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *CVPR*, 2023.

[17] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," in *ICLR*, 2022.

[18] V. Goel, E. Peruzzo, Y. Jiang, D. Xu, X. Xu, N. Sebe, T. Darrell, Z. Wang, and H. Shi, "Pair-diffusion: A comprehensive multimodal object-level image editor," *arXiv preprint*, 2023.

[19] O. Avrahami, T. Hayes, O. Gafni, S. Gupta, Y. Taigman, D. Parikh, D. Lischinski, O. Fried, and X. Yin, "Spatext: Spatio-textual representation for controllable image generation," in *CVPR*, 2023.

[20] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023.

[21] L. Huang, D. Chen, Y. Liu, Y. Shen, D. Zhao, and J. Zhou, "Composer: Creative and controllable image synthesis with composable conditions," in *ICML*, 2023.

[22] J. Cho, A. Zala, and M. Bansal, "Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models," in *ICCV*, 2023.

[23] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *NeurIPS*, 2016.

[24] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, "Women also snowboard: Overcoming bias in captioning models," in *ECCV*, 2018.

[25] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," in *EMNLP*, 2017.

[26] R. Naik and B. Nushi, "Social biases through the text-to-image generation lens," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023.

[27] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the international workshop on software fairness*, 2018.

[28] D. Xu, S. Yuan, L. Zhang, and X. Wu, "Fairgan: Fairness-aware generative adversarial networks," in *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, 2018.

[29] J. Nam, H. Cha, S. Ahn, J. Lee, and J. Shin, "Learning from failure: De-biasing classifier from biased classifier," *NeurIPS*, 2020.

[30] Y. Savani, C. White, and N. S. Govindarajulu, "Intra-processing methods for debiasing neural networks," in *NeurIPS*, 2020.

[31] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky, "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *CVPR*, 2020.

[32] S. Jung, S. Chun, and T. Moon, "Learning fair classifiers with partially annotated group labels," in *CVPR*, 2022.

[33] S. Agarwal, S. Muku, S. Anand, and C. Arora, "Does data repair lead to fair models? curating contextually fair data to reduce model bias," in *WACV*, 2022.

[34] M. D'Inca, C. Tzelepis, Y. Patras, and N. Sebe, "Improving fairness using vision-language driven image augmentation," in *WACV*, 2024.

[35] X. Su, Y. Ren, W. Qiang, Z. Song, H. Gao, F. Wu, and C. Zheng, "Unbiased image synthesis via manifold-driven sampling in diffusion models," *arXiv preprint*, 2023.

[36] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, 2014.

[37] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *ECCV*, 2014.

[38] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *WACV*, 2021.

[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[40] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *ICML*, 2022.

[41] C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao, *et al.*, "mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

[42] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.

[43] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.

[44] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[45] M. D'Inca, E. Peruzzo, M. Mancini, D. Xu, V. Goel, X. Xu, Z. Wang, H. Shi, and N. Sebe, "Openbias: Open-set bias detection in generative models," in *CVPR*, 2024.

[46] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. The MIT Press, 2019.

[47] N. Smuha, "Ethics guidelines for trustworthy AI," in *AI & Ethics, Date: 2019/05/28-2019/05/28, Brussels, Belgium*, European Commission, 2019.

[48] N. Oliver, "Artificial intelligence for social good - The way forward," in *Science, Research and Innovation performance of the EU 2022 report*, ch. 11, pp. 604–707, European Commission, 2022.

[49] A. N. Carey and X. Wu, "The statistical fairness field guide: perspectives from social and formal sciences," *AI and Ethics*, pp. 1–23, 2022.

[50] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[51] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness Beyond Disparate Treatment &; Disparate Impact: Learning Classification without Disparate Mistreatment," in *International Conference on World Wide Web*, p. 1171–1180, 2017.

[52] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

[53] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[54] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[55] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, pp. 325–333, PMLR, 2013.

[56] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

[57] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Joint European conference on machine learning and knowledge discovery in databases*, pp. 35–50, Springer, 2012.

[58] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.

[59] P. Hacker and J.-H. Passoth, "Varieties of ai explanations under the law. from the gdpr to the aia, and beyond," in *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pp. 343–373, Springer, 2022.

[60] M. Madaio, L. Egede, H. Subramonyam, J. Wortman Vaughan, and H. Wallach, "Assessing the fairness of ai systems: Ai practitioners' processes, challenges, and needs for support," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW1, pp. 1–26, 2022.

[61] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[62] T. Hagendorff, "The ethics of ai ethics: An evaluation of guidelines," *Minds and Machines*, vol. 30, no. 1, pp. 99–120, 2020.

[63] A. Chouldechova and A. Roth, "A snapshot of the frontiers of fairness in machine learning," *Communications of the ACM*, vol. 63, no. 5, pp. 82–89, 2020.

[64] A. Ghorbani and J. Zou, "Data shapley: Equitable valuation of data for machine learning," in *International Conference on Machine Learning*, pp. 2242–2251, PMLR, 2019.

[65] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, pp. 307–317, 1953.

[66] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. M. Gurel, B. Li, C. Zhang, C. Spanos, and D. Song, "Efficient task-specific data valuation for nearest neighbor algorithms," *Proc. VLDB Endow.*, vol. 12, no. 11, p. 1610–1623, 2019.

[67] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44, 2021.

[68] J. Chai and X. Wang, "Fairness with adaptive weights," in *International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research*, pp. 2853–2866, PMLR, 17–23 Jul 2022.

[69] F. Kamiran and T. Calders, "Classifying without discriminating," in *2009 2nd international conference on computer, control and communication*, pp. 1–6, IEEE, 2009.

[70] R. Kohavi *et al.*, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.," in *Kdd*, vol. 96, pp. 202–207, 1996.

[71] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. propublica, may 23," 2016.

[72] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[73] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," in *International Conference on Artificial Intelligence and Statistics*, pp. 702–712, PMLR, 2020.

[74] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[75] P. Li and H. Liu, "Achieving fairness at no utility cost via data reweighing with influence," in *International Conference on Machine Learning*, vol. 162, pp. 12917–12930, PMLR, 17–23 Jul 2022.

[76] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

[77] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1548–1558, 2021.

[78] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[79] E. Albini, J. Long, D. Dervovic, and D. Magazzeni, "Counterfactual shapley additive explanations," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1054–1070, 2022.

[80] A. Arnaiz-Rodriguez and N. Oliver, "Towards algorithmic fairness by means of instance-level data re-weighting based on shapley values," in *ICLR 2024 Workshop on Data-centric Machine Learning Research (DMLR): Harnessing Momentum for Science*, 2024.

[81] P. Bedapudi, "NudeNet: Neural Nets for Nudity Classification, Detection, and Selective Censoring," 2019.

[82] D. L. Crone, S. Bode, C. Murawski, and S. M. Laham, "The Socio-Moral Image Database (SMID): A novel stimulus set for the study of social, moral and affective processes," *PloS one*, vol. 13, no. 1, p. e0190954, 2018.

[83] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, *et al.*, "DataComp: In search of the next generation of multimodal datasets," in *NeurIPS*, 2024.

[84] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs," 2021.

[85] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *CVPR*, 2022.

[86] P. Schramowski, C. Tauchmann, and K. Kersting, "Can Machines Help Us Answering Question 16 in Datasheets, and In Turn Reflecting on Inappropriate Content?," in *ACM FAccT*, 2022.

[87] D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, L. Baraldi, M. Cornia, and R. Cucchiara, "The Revolution of Multimodal Large Language Models: A Survey," in *ACL Findings*, 2024.

[88] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," in *NeurIPS*, 2023.

[89] S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, and R. Cucchiara, "Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models," in *Proceedings of the European Conference on Computer Vision*, 2024.

[90] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep Long-Tailed Learning: A Survey," 2023.

[91] D. Samuel, R. Ben-Ari, S. Raviv, N. Darshan, and G. Chechik, "Generating images of rare concepts using pre-trained diffusion models," in *AAAI*, 2023.

[92] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models," *ACM Transactions on Graphics (TOG)*, 2023.

[93] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum, "Compositional Visual Generation with Composable Diffusion Models," in *ECCV*, 2022.

[94] W. Feng, X. He, T.-J. Fu, V. Jampani, A. R. Akula, P. Narayana, S. Basu, X. E. Wang, and W. Y. Wang, "Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis," in *ICLR*, 2023.

[95] A. Helbling, E. Montoya, and D. H. Chau, "ObjectComposer: Consistent Generation of Multiple Objects Without Fine-tuning," *arXiv preprint arXiv:2310.06968*, 2023.

[96] S. Karthik, K. Roth, M. Mancini, and Z. Akata, "If at First You Don't Succeed, Try, Try Again: Faithful Diffusion-based Text-to-Image Generation by Selection," *arXiv preprint arXiv:2305.13308*, 2023.

[97] D. Samuel, R. Ben-Ari, N. Darshan, H. Maron, and G. Chechik, "Norm-guided latent space exploration for text-to-image generation," in *NeurIPS*, 2024.

[98] Q. Wu, Y. Liu, H. Zhao, T. Bui, Z. Lin, Y. Zhang, and S. Chang, "Harnessing the Spatial-Temporal Attention of Diffusion Models for High-Fidelity Text-to-Image Synthesis," in *ICCV*, 2023.

[99] F. Betti, J. Staiano, L. Baraldi, L. Baraldi, R. Cucchiara, and N. Sebe, "Let's vice! mimicking human cognitive behavior in image generation evaluation," in *ACM MM*, 2023.

[100] Y. Lu, X. Yang, X. Li, X. E. Wang, and W. Y. Wang, "LLMScore: Unveiling the Power of Large Language Models in Text-to-Image Synthesis Evaluation," in *NeurIPS*, 2024.

[101] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, pp. 5998–6008, 2017.

[102] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum, "Compositional Visual Generation with Composable Diffusion Models," in *ECCV*, 2022.

[103] M. Chen, I. Laina, and A. Vedaldi, "Training-Free Layout Control with Cross-Attention Guidance," in *WACV*, 2024.

[104] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling Open-Vocabulary Object Detection," in *NeurIPS*, 2024.

[105] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[106] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances*, vol. 4, no. 1, p. eaao5580, 2018.

[107] M. A. Sheikh, A. K. Goel, and T. Kumar, "An approach for prediction of loan approval using machine learning algorithm," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 490–494, IEEE, 2020.

[108] C. Liem, M. Langer, A. Demetriou, A. M. Hiemstra, A. Sukma Wicaksana, M. P. Born, and C. J. König, "Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening," in *Explainable and interpretable models in computer vision and machine learning*, pp. 197–253, Springer, 2018.

[109] T. K. Yoo, I. H. Ryu, G. Lee, Y. Kim, J. K. Kim, I. S. Lee, J. S. Kim, and T. H. Rim, "Adopting machine learning to automatically identify candidate patients for corneal refractive surgery," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–9, 2019.

[110] S. Venkatasubramanian and M. Alfano, "The philosophical basis of algorithmic recourse," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 284–293, 2020.

[111] A.-H. Karimi, B. Schölkopf, and I. Valera, "Algorithmic recourse: from counterfactual explanations to interventions," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 353–362, 2021.

[112] G. De Toni, B. Lepri, and A. Passerini, "Synthesizing explainable counterfactual policies for algorithmic recourse with program synthesis," *Mach. Learn.*, vol. 112, p. 1389–1409, feb 2023.

[113] P. Naumann and E. Ntoutsi, "Consequence-aware sequential counterfactual generation," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 682–698, Springer, 2021.

[114] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera, "A survey of algorithmic recourse: Contrastive explanations and consequential recommendations," *ACM Comput. Surv.*, vol. 55, dec 2022.

[115] G. Ramakrishnan, Y. C. Lee, and A. Albarghouthi, "Synthesizing action sequences for modifying model decisions," in *AAAI*, vol. 34, pp. 5462–5469, 2020.

[116] S. Yonadav and W. S. Moses, "Extracting incentives from black-box decisions," *CoRR*, vol. abs/1910.05664, 2019.

[117] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *FAT**, pp. 10–19, 2019.

[118] A.-H. Karimi, J. Von Kügelgen, B. Schölkopf, and I. Valera, "Algorithmic recourse under imperfect causal knowledge: a probabilistic approach," *Advances in Neural Information Processing Systems*, vol. 33, pp. 265–277, 2020.

[119] S. Tsirtsis and M. Rodriguez, "Decisions, counterfactual explanations and strategic behavior," in *NeurIPS*, 2020.

[120] C. Russell, "Efficient search for diverse coherent explanations," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, (New York, NY, USA), p. 20–28, Association for Computing Machinery, 2019.

[121] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *FAT**, pp. 607–617, 2020.

[122] Z. J. Wang, J. Wortman Vaughan, R. Caruana, and D. H. Chau, "Gam coach: Towards interactive and user-centered algorithmic recourse," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2023.

[123] S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-objective counterfactual explanations," in *PPSN*, pp. 448–469, Springer, 2020.

[124] P. Yadav, P. Hase, and M. Bansal, "Low-cost algorithmic recourse for users with uncertain cost functions," *arXiv preprint arXiv:2111.01235*, 2021.

[125] K. Rawal and H. Lakkaraju, "Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12187–12198, 2020.

[126] D. Mahajan, C. Tan, and A. Sharma, "Preserving causal constraints in counterfactual explanations for machine learning classifiers," *arXiv preprint arXiv:1912.03277*, 2019.

[127] U. Chajewska, D. Koller, and R. Parr, "Making rational decisions using adaptive utility elicitation," in *AAAI/IAAI*, pp. 363–369, 2000.

[128] C. Boutilier, "A pomdp formulation of preference elicitation problems," in *AAAI*, pp. 239–246, 2002.

[129] D. Braziunas and C. Boutilier, "Minimax regret based elicitation of generalized additive utilities," in *UAI*, pp. 25–32, 2007.

[130] S. Guo and S. Sanner, "Real-time multiattribute bayesian preference elicitation with pairwise comparison queries," in *AISTATS*, pp. 289–296, 2010.

[131] P. Viappiani and C. Boutilier, "Optimal Bayesian recommendation sets and myopically optimal choice query sets," *Advances in neural information processing systems*, vol. 23, 2010.

[132] P. Dragone, S. Teso, and A. Passerini, "Constructive preference elicitation," *Frontiers in Robotics and AI*, vol. 4, p. 71, 2018.

[133] P. Viappiani and C. Boutilier, "On the equivalence of optimal recommendation sets and myopically optimal query sets," *Artificial Intelligence*, vol. 286, p. 103328, 2020.

[134] R. Price and P. R. Messinger, "Optimal recommendation sets: Covering uncertainty over user preferences," in *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA* (M. M. Veloso and S. Kambhampati, eds.), pp. 541–548, AAAI Press / The MIT Press, 2005.

[135] R. Akrour, M. Schoenauer, and M. Sebag, "APRIL: active preference learning-based reinforcement learning," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II* (P. A. Flach, T. D. Bie, and N. Cristianini, eds.), vol. 7524 of *Lecture Notes in Computer Science*, pp. 116–131, Springer, 2012.

[136] R. D. Luce, *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2012.

[137] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—i," *Mathematical programming*, vol. 14, no. 1, pp. 265–294, 1978.

[138] A. Krause and D. Golovin, "Submodular function maximization.," *Tractability*, vol. 3, pp. 71–104, 2014.

[139] Kaggle, "Give me some credit," Sep 2011.

[140] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[141] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "Face: feasible and actionable counterfactual explanations," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 344–350, 2020.

[142] J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, T. Pfaff, T. Weber, L. Buesing, and P. W. Battaglia, "Combining q-learning and search with amortized value estimates," *arXiv preprint arXiv:1912.02807*, 2019.

[143] S. Verma, K. Hines, and J. P. Dickerson, "Amortized generation of sequential algorithmic recourses for black-box models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 8512–8519, 2022.

[144] S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, and C. Shah, "Counterfactual explanations and algorithmic recourses for machine learning: A review," *arXiv preprint arXiv:2010.10596*, 2020.

[145] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[146] P. Voigt and A. Von dem Bussche, "The EU general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, no. 3152676, pp. 10–5555, 2017.

[147] G. D. Toni, P. Viappiani, S. Teso, B. Lepri, and A. Passerini, "Personalized algorithmic recourse with preference elicitation," *Transactions on Machine Learning Research*, 2024.

[148] S. Esfahani, G. De Toni, B. Lepri, A. Passerini, K. Tentori, and M. Zancanaro, "Preference elicitation in interactive and user-centered algorithmic recourse: an initial exploration," in *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pp. 249–254, 2024.

[149] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, *et al.*, "Bias in data-driven artificial intelligence systems—an introductory survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1356, 2020.

[150] S. Caton and C. Haas, "Fairness in machine learning: A survey," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–38, 2024.

[151] M. S. A. Lee and J. Singh, "The landscape and gaps in open source fairness toolkits," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–13, 2021.

[152] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, *et al.*, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[153] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A toolkit for assessing and improving fairness in ai," *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[154] L. Cheng, K. R. Varshney, and H. Liu, "Socially responsible ai algorithms: Issues, purposes, and challenges," *Journal of Artificial Intelligence Research*, vol. 71, pp. 1137–1181, 2021.

[155] L. Weinberg, "Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ml fairness approaches," *Journal of Artificial Intelligence Research*, vol. 74, pp. 75–109, 2022.

[156] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *8th Innovations in Theoretical Computer Science*, 2017.

[157] E. Krasanakis and S. Papadopoulos, "Evaluating ai group fairness: a fuzzy logic perspective," *arXiv preprint arXiv:2406.18939*, 2024.

[158] L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy sets and systems*, vol. 1, no. 1, pp. 3–28, 1978.

[159] P. Hájek, "Basic fuzzy logic and bl-algebras," *Soft computing*, vol. 2, pp. 124–128, 1998.

[160] R. Cignoli, F. Esteva, L. Godo, and A. Torrens, "Basic fuzzy logic is the logic of continuous t-norms and their residua," *Soft computing*, vol. 4, pp. 106–112, 2000.

[161] F. Esteva and L. Godo, "Monoidal t-norm based logic: towards a logic for left-continuous t-norms," *Fuzzy sets and systems*, vol. 124, no. 3, pp. 271–288, 2001.

[162] L. Spada and A. Di Nola, "A short introduction to formal fuzzy logic via t-norms," in *Metodi, Modelli eTecnologie dell'informazione a Supporto delle Decisioni*, vol. 1, pp. 194–202, Franco Angeli, 2008.

[163] P. Cintula, C. G. Fermüller, and C. Noguera, "Fuzzy Logic," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta and U. Nodelman, eds.), Metaphysics Research Lab, Stanford University, Summer 2023 ed., 2023.

[164] J. M. Rohrer, "Thinking clearly about correlations and causation: Graphical causal models for observational data," *Advances in methods and practices in psychological science*, vol. 1, no. 1, pp. 27–42, 2018.

[165] H. M. Blalock Jr, *Causal models in the social sciences*. Routledge, 2017.

[166] E. P. Klement, R. Mesiar, and E. Pap, *Triangular norms*, vol. 8. Springer Science & Business Media, 2013.

[167] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan, "An intersectional definition of fairness," in *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 1918–1921, IEEE, 2020.

[168] D. Biddle, *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Routledge, 2017.

[169] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data mining and knowledge discovery*, vol. 21, pp. 277–292, 2010.

[170] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica*, vol. 47, p. 263, Mar. 1979.

[171] H. A. Simon, "A Behavioral Model of Rational Choice," *The Quarterly Journal of Economics*, vol. 69, pp. 99–118, 02 1955.

[172] D. Kahneman, *Thinking, fast and slow*. Macmillan, 2011.

[173] D. Ariely and S. Jones, *Predictably irrational*. 2008.

[174] N. D. Schüll, *Addiction by design*. Princeton University Press, 2012.

[175] N. Eyal, *Hooked: How to build habit-forming products*. Penguin, 2014.

[176] M. Petticrew, N. Maani, L. Pettigrew, H. Rutter, and M. C. Van Schalkwyk, "Dark nudges and sludge in big alcohol: Behavioral economics, cognitive biases, and alcohol industry corporate social responsibility," *The Milbank Quarterly*, vol. 98, pp. 1290–1328, Sept. 2020.

[177] L. M. Hiatt, C. Narber, E. Bekele, S. S. Khemlani, and J. G. Trafton, "Human modeling for human–robot collaboration," *The International Journal of Robotics Research*, Feb. 2017.

[178] A. S. Rich and T. M. Gureckis, "Lessons for artificial intelligence from the study of natural stupidity," *Nature Machine Intelligence*, vol. 1, no. 4, pp. 174–180, 2019.

[179] C. Rastogi, Y. Zhang, D. Wei, K. R. Varshney, A. Dhurandhar, and R. Tomsett, "Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making," 2020.

[180] T. Kliegr, Š. Bahník, and J. Fürnkranz, "A review of possible effects of cognitive biases on interpretation of rule-based machine learning models," *Artificial Intelligence*, vol. 295, June 2021.

[181] J. S. Blumenthal-Barby and H. Krieger, "Cognitive biases and heuristics in medical decision making," *Medical Decision Making*, vol. 35, pp. 539–557, Aug. 2014.

[182] G. Saposnik, D. Redelmeier, C. C. Ruff, *et al.*, "Cognitive biases associated with medical decisions: a systematic review," *BMC Medical Informatics and Decision Making*, Nov. 2016.

[183] W. Wattanacharoensil and D. La-ornual, "A systematic review of cognitive biases in tourist decisions," *Tourism Management*, vol. 75, pp. 353–369, Dec. 2019.

[184] M. J. Kinsey, S. M. V. Gwynne, E. D. Kuligowski, and M. Kinateder, "Cognitive biases within decision making during fire evacuations," *Fire Technology*, vol. 55, pp. 465–485, Mar. 2018.

[185] E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic, "A task-based taxonomy of cognitive biases for information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, pp. 1413–1432, Feb. 2020.

[186] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty.," *Science*, vol. 185, p. 1124–1131, Sept. 1974.

[187] K. E. Stanovich, M. E. Toplak, and R. F. West, "The development of rational thought: A taxonomy of heuristics and biases," in *Advances in Child Development and Behavior*, pp. 251–285, Elsevier, 2008.

[188] D. Arnott, "Cognitive biases and decision support systems development: a design science approach," *Information Systems Journal*, vol. 16, Jan. 2006.

[189] R. Pohl and R. F. Pohl, *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press, 2004.

[190] J. Huber, J. W. Payne, and C. Puto, "Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis," *Journal of Consumer Research*, vol. 9, p. 90, June 1982.

[191] J. Hu and R. Yu, "The neural correlates of the decoy effect in decisions," *Frontiers in Behavioral Neuroscience*, vol. 8, Aug. 2014.

[192] Z. Wang, M. Jusup, L. Shi, *et al.*, "Exploiting a cognitive bias promotes cooperation in social dilemma experiments," *Nature Communications*, vol. 9, July 2018.

[193] B. M. Josiam and J. P. Hobson, "Consumer choice in context: The decoy effect in travel and tourism," *Journal of Travel Research*, vol. 34, pp. 45–50, July 1995.

[194] E. C. Teppan and A. Felfernig, "Minimization of decoy effects in recommender result sets," *Web Intelligence and Agent Systems: An International Journal*, vol. 10, no. 4, pp. 385–395, 2012.

[195] A. Tversky and D. Kahneman, "The framing of decisions and the psychology of choice," *Science*, vol. 211, no. 4481, 1981.

[196] S. Gächter, H. Orzen, E. Renner, and C. Starmer, "Are experimental economists prone to framing effects? a natural field experiment," *Journal of Economic Behavior & Organization*, vol. 70, pp. 443–446, June 2009.

[197] I. P. Levin, S. L. Schneider, and G. J. Gaeth, "All frames are not created equal: A typology and critical analysis of framing effects," *Organizational Behavior and Human Decision Processes*, vol. 76, pp. 149–188, Nov. 1998.

[198] J. Gong, Y. Zhang, Z. Yang, Y. Huang, J. Feng, and W. Zhang, "The framing effect in medical decision-making: a review of the literature," *Psychology, Health & Medicine*, vol. 18, pp. 645–653, Dec. 2013.

[199] P. E. Souza, C. P. C. Chanel, F. Dehais, and S. Givigi, "Towards human-robot interaction: A framing effect experiment," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 001929–001934, IEEE, Oct. 2016.

[200] T. Kim and H. Song, "Communicating the limitations of AI: The effect of message framing and ownership on trust in artificial intelligence," *International Journal of Human–Computer Interaction*, pp. 1–11, Apr. 2022.

[201] F. Ni, D. Arnott, and S. Gao, "The anchoring effect in business intelligence supported decision-making," *Journal of Decision Systems*, vol. 28, pp. 67–81, Apr. 2019.

[202] T. Yasseri and J. Reher, "Fooled by facts: quantifying anchoring bias through a large-scale experiment," *Journal of Computational Social Science*, vol. 5, pp. 1001–1021, Jan. 2022.

[203] G. Adomavicius, J. C. Bockstedt, S. P. Curley, and J. Zhang, "Do recommender systems manipulate consumer preferences? a study of anchoring effects," *Inf. Syst. Res.*, Dec. 2013.

[204] B. K. Hayes and B. R. Newell, "Induction with uncertain categories: When do people consider the category alternatives?," *Memory & Cognition*, vol. 37, pp. 730–743, Sept. 2009.

[205] D. V. Burakov, "Exogenous credit cycle: An experimental study," *World Applied Sciences Journal*, vol. 26, no. 6, 2013.

[206] A. Tversky and D. Kahneman, "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment.," *Psychological Review*, vol. 90, no. 4, pp. 293–315, 1983.

[207] K. Tentori, N. Bonini, and D. Osherson, "The conjunction fallacy: a misunderstanding about conjunction?," *Cognitive Science*, vol. 28, pp. 467–477, May 2004.

[208] D. H. Wedell and R. Moro, "Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus, and problem type," *Cognition*, vol. 107, Apr. 2008.

[209] Y. Lo, A. Sides, J. Rozelle, and D. Osherson, "Evidential diversity and premise probability in young children's inductive judgment," *Cognitive Science*, vol. 26, pp. 181–206, Mar. 2002.

[210] A. K. Barbey and S. A. Sloman, "Base-rate respect: From ecological rationality to dual processes," *Behavioral and Brain Sciences*, vol. 30, pp. 241–254, June 2007.

[211] M. Bar-Hillel, "The base-rate fallacy in probability judgments," *Acta Psychologica*, vol. 44, pp. 211–233, May 1980.

[212] E. Gold, *The gambler's fallacy*. PhD thesis, Carnegie Mellon University, 1997.

[213] G. Barron and S. Leider, "The role of experience in the gambler's fallacy," *Journal of Behavioral Decision Making*, vol. 23, pp. 117–129, Jan. 2010.

[214] D. L. Chen, T. J. Moskowitz, and K. Shue, "Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires," *The Quarterly Journal of Economics*, vol. 131, pp. 1181–1242, Mar. 2016.

[215] R. Thaler, "Some empirical evidence on dynamic inconsistency," *Economics Letters*, vol. 8, pp. 201–207, Jan. 1981.

[216] W. H. Hampton, N. Asadi, and I. R. Olson, "Good things for those who wait: Predictive modeling highlights importance of delay discounting for income attainment," *Frontiers in Psychology*, vol. 9, Sept. 2018.

[217] G. Ainslie, "Specious reward: A behavioral theory of impulsiveness and impulse control.," *Psychological Bulletin*, 1975.

[218] C. F. Kurz and A. N. König, "Predicting time preference from social media behavior," *Future Generation Computer Systems*, vol. 130, pp. 155–163, May 2022.

[219] T. S. van Endert and P. N. C. Mohr, "Delay discounting of monetary and social media rewards: Magnitude and trait effects," *Frontiers in Psychology*, vol. 13, Feb. 2022.

[220] K. Dion, E. Berscheid, and E. Walster, "What is beautiful is good.," *Journal of Personality and Social Psychology*, vol. 24, no. 3, pp. 285–290, 1972.

[221] R. E. Nisbett and T. D. Wilson, "The halo effect: Evidence for unconscious alteration of judgments.," *Journal of Personality and Social Psychology*, vol. 35, pp. 250–256, Apr. 1977.

[222] J. L. Gibson and J. S. Gore, "Is he a hero or a weirdo? how norm violations influence the halo effect," *Gender Issues*, vol. 33, pp. 299–310, Sept. 2016.

[223] D. Landy and H. Sigall, "Beauty is talent: Task evaluation as a function of the performer's physical attractiveness.," *Journal of Personality and Social Psychology*, vol. 29, no. 3, 1974.

[224] M. I. Norton, D. Mochon, and D. Ariely, "The IKEA effect: When labor leads to love," *Journal of Consumer Psychology*, vol. 22, pp. 453–460, July 2012.

[225] T. Radtke, N. Liszewska, K. Horodyska, M. Boberska, K. Schenkel, and A. Luszczynska, "Cooking together: The ikea effect on family vegetable intake," *British Journal of Health Psychology*, vol. 24, pp. 896–912, Sept. 2019.

[226] F. Brunner, F. Gamm, and W. Mill, "MyPortfolio: The IKEA effect in financial investment decisions," *Journal of Banking & Finance*, p. 106529, May 2022.

[227] J. W. Pratt, "Risk aversion in the small and in the large," in *Uncertainty in Economics*, pp. 59–79, Elsevier, 1978.

[228] K. E. Stanovich, *Decision making and rationality in the modern world*. New York, Oxford University Press, 2010.

[229] B. Fischhoff, P. Slovic, S. Lichtenstein, *et al.*, "How safe is safe enough? a psychometric study of attitudes towards technological risks and benefits," *Policy Sciences*, Apr. 1978.

[230] Y. Rottenstreich and C. K. Hsee, "Money, kisses, and electric shocks: On the affective psychology of risk," *Psychological Science*, vol. 12, pp. 185–190, May 2001.

[231] D. P. Crowne and D. Marlowe, "A new scale of social desirability independent of psychopathology.," *Journal of Consulting Psychology*, vol. 24, no. 4, pp. 349–354, 1960.

[232] J. R. Hebert, L. Clemow, L. Pbert, I. S. Ockene, and J. K. Ockene, "Social desirability bias in dietary self-report may compromise the validity of dietary intake measures," *International Journal of Epidemiology*, vol. 24, no. 2, 1995.

[233] L. Harrison, "The validity of self-reported drug use in survey research: An overview and critique of research methods. national institute of drug abuse monograph 167," 2006.

[234] G. S. Stuart and D. A. Grimes, "Social desirability bias in family planning studies: a neglected problem," *Contraception*, vol. 80, pp. 108–112, Aug. 2009.

[235] E. F. Loftus and J. C. Palmer, "Reconstruction of automobile destruction: An example of the interaction between language and memory," *Journal of Verbal Learning and Verbal Behavior*, vol. 13, pp. 585–589, Oct. 1974.

[236] E. F. Loftus, "Reconstructing memory: The incredible eyewitness," *Jurimetrics Journal*, vol. 15, no. 3, pp. 188–193, 1975.

[237] J. M. Lampinen, J. S. Neuschatz, and D. G. Payne, "Memory illusions and consciousness: Examining the phenomenology of true and false memories," *Current Psychology*, vol. 16, pp. 181–224, Sept. 1997.

[238] D. M. Bernstein and E. F. Loftus, "The consequences of false memories for food preferences and choices," *Perspectives on Psychological Science*, vol. 4, pp. 135–139, Mar. 2009.

[239] T. B. Rogers, N. A. Kuiper, and W. S. Kirker, "Self-reference and the encoding of personal information.," *Journal of Personality and Social Psychology*, vol. 35, no. 9, pp. 677–688, 1977.

[240] A. H. Gutchess, E. A. Kensinger, C. Yoon, and D. L. Schacter, "Ageing and the self-reference effect in memory," *Memory*, vol. 15, pp. 822–837, Nov. 2007.

[241] B. B. Murdock, "The serial position effect of free recall.," *Journal of Experimental Psychology*, vol. 64, pp. 482–488, Nov. 1962.

[242] B. Murdock and J. Metcalfe, "Controlled rehearsal in single-trial free recall," *Journal of Verbal Learning and Verbal Behavior*, vol. 17, pp. 309–324, June 1978.

[243] S. E. Asch, "Forming impressions of personality.," *The Journal of Abnormal and Social Psychology*, vol. 41, July 1946.

[244] D. Kahneman, B. L. Fredrickson, C. A. Schreiber, and D. A. Redelmeier, "When more pain is preferred to less: Adding a better end," *Psychological Science*, vol. 4, pp. 401–405, Nov. 1993.

[245] Z. Carmon and D. Kahneman, "The experienced utility of queuing: real time affect and retrospective evaluations of simulated queues," *Duke University: Durham, NC, USA*, 1996.

[246] P. De Maeyer and H. Estelami, "Applying the peak-end rule to reference prices," *Journal of Product & Brand Management*, 2013.

[247] W. Samuelson and R. Zeckhauser, "Status quo bias in decision making," *Journal of Risk and Uncertainty*, vol. 1, Mar. 1988.

[248] D. Kahneman, J. L. Knetsch, and R. H. Thaler, "Anomalies: The endowment effect, loss aversion, and status quo bias," *Journal of Economic Perspectives*, vol. 5, Feb. 1991.

[249] D. R. Forsyth, "Group dynamics," 1990.

[250] T. Postmes, R. Spears, and S. Cihangir, "Quality of decision making and group norms.," *Journal of Personality and Social Psychology*, vol. 80, no. 6, pp. 918–930, 2001.

[251] G. Stasser and D. Stewart, "Discovery of hidden profiles by decision-making groups: Solving a problem versus making a judgment.," *Journal of Personality and Social Psychology*, vol. 63, pp. 426–434, Sept. 1992.

[252] I. Simonson, "The effect of purchase quantity and timing on variety-seeking behavior," *Journal of Marketing Research*, vol. 27, p. 150, May 1990.

[253] D. Read and G. Loewenstein, "Diversification bias: Explaining the discrepancy in variety seeking between combined and separated choices.," *Journal of Experimental Psychology: Applied*, vol. 1, pp. 34–49, Mar. 1995.

[254] D. Kliger, M. J. van den Assem, and R. C. Zwinkels, "Empirical behavioral finance," *Journal of Economic Behavior & Organization*, vol. 107, pp. 421–427, Nov. 2014.

[255] G. Gigerenzer and D. G. Goldstein, "Reasoning the fast and frugal way: Models of bounded rationality.," *Psychological Review*, vol. 103, pp. 650–669, Oct. 1996.

[256] Y. Wang, S. Luan, and G. Gigerenzer, "Modeling fast-and-frugal heuristics," *PsyCh Journal*, vol. 11, pp. 600–611, July 2022.

[257] C. A. Hidalgo *et al.*, *How humans judge machines*. MIT Press, 2021.

[258] H. Taniguchi, H. Sato, and T. Shirakawa, "Application of human cognitive mechanisms to naïve bayes text classifier," *AIP Conference Proceedings*, vol. 1863, no. 1, p. 360016, 2017.

[259] H. Taniguchi, H. Sato, and T. Shirakawa, "Implementation of human cognitive bias on neural network and its application to breast cancer diagnosis," *SICE Journal of Control, Measurement, and System Integration*, vol. 12, Mar. 2019.

[260] M. Sidman, R. Rauzin, R. Lazar, S. Cunningham, W. Tailby, and P. Carrigan, "A search for symmetry in the conditional discriminations of rhesus monkeys, baboons, and children," *Journal of the Experimental Analysis of Behavior*, vol. 37, pp. 23–44, Jan. 1982.

[261] W. E. Merriman, L. L. Bowman, and B. MacWhinney, "The mutual exclusivity bias in children's word learning," *Monographs of the Society for Research in Child Development*, vol. 54, no. 3/4, p. i, 1989.

[262] R. S. Crowley, E. Legowski, O. Medvedeva, *et al.*, "Automated detection of heuristics and biases among pathologists in a computer-based system," *Advances in Health Sciences Education*, vol. 18, pp. 343–363, May 2012.

[263] M. McShane, S. Nirenburg, and B. Jarrell, "Modeling decision-making biases," *Biologically Inspired Cognitive Architectures*, vol. 3, pp. 39–50, Jan. 2013.

[264] N. J. Blunch, "Position bias in multiple-choice questions," *Journal of Marketing Research*, vol. 21, pp. 216–220, May 1984.

[265] J.-H. Kang and K. Lerman, "VIP: Incorporating human cognitive biases in a probabilistic model of retweeting," in *Social Computing, Behavioral-Cultural Modeling, and Prediction*, pp. 101–110, Springer International Publishing, 2015.

[266] J. Tenenbaum, "Bayesian modeling of human concept learning," *Advances in neural information processing systems*, vol. 11, 1998.

[267] T. L. Griffiths and J. B. Tenenbaum, "Optimal predictions in everyday cognition," *Psychological Science*, vol. 17, Sept. 2006.

[268] N. Chater, J. B. Tenenbaum, and A. Yuille, "Probabilistic models of cognition: Conceptual foundations," *Trends in Cognitive Sciences*, vol. 10, pp. 287–291, July 2006.

[269] R. A. Jansen, A. N. Rafferty, and T. L. Griffiths, "A rational model of the dunning–kruger effect supports insensitivity to evidence in low performers," *Nature Human Behaviour*, vol. 5, pp. 756–763, Feb. 2021.

[270] A. Gulati, M. A. Lozano, B. Lepri, and N. Oliver, "Biased: Bringing irrationality into automated system design," 2023.

[271] S. N. Talamas, *Perceptions of intelligence and the attractiveness halo*. PhD thesis, University of St Andrews, 2016.

[272] G. Perrotta, "The concept of altered perception in "body dysmorphic disorder": the subtle border between the abuse of selfies in social networks and cosmetic surgery, between socially accepted dysfunctionality and the pathological condition," *Journal of Neurology, Neurological Science and Disorders*, vol. 6, no. 1, pp. 001–007, 2020.

[273] R. T. Cristel, S. H. Dayan, M. Akinosun, and P. T. Russell, "Evaluation of selfies and filtered selfies and effects on first impressions," *Aesthetic Surgery Journal*, vol. 41, no. 1, pp. 122–130, 2021.

[274] P. Riccio, B. Psomas, F. Galati, F. Escolano, T. Hofmann, and N. Oliver, "Openfilter: a framework to democratize research access to social media ar filters," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12491–12503, 2022.

[275] P. Riccio and N. Oliver, "Racial bias in the beautyverse: Evaluation of augmented-reality beauty filters," in *European Conference on Computer Vision*, pp. 714–721, Springer, 2022.

[276] D. S. Ma, J. Correll, and B. Wittenbrink, "The chicago face database: A free stimulus set of faces and norming data," *Behavior Research Methods*, vol. 47, pp. 1122–1135, Jan. 2015.

[277] N. C. Ebner, M. Riediger, and U. Lindenberger, "FACES—a database of facial expressions in young, middle-aged, and older women and men: Development and validation," *Behavior Research Methods*, vol. 42, pp. 351–362, Feb. 2010.

[278] R. Barocas and P. Karoly, "Effects of physical appearance on social responsiveness," *Psychological Reports*, vol. 31, pp. 495–500, Oct. 1972.

[279] C. Batres and V. Shiramizu, "Examining the "attractiveness halo effect" across cultures," *Current Psychology*, Aug. 2022.

[280] B. J. Guise, C. H. Pollans, and I. D. Turkat, "Effects of physical attractiveness on perception of social skill," *Perceptual and Motor Skills*, vol. 54, pp. 1039–1042, June 1982.

[281] J. R. Kunst, J. Kirkøen, and O. Mohamdain, "Hacking attractiveness biases in hiring? the role of beautifying photo-filters," *Management Decision*, vol. 61, pp. 924–943, Apr. 2022.

[282] K. N. Lewis and W. B. Walsh, "Physical attractiveness: Its impact on the perception of a female counselor.," *Journal of Counseling Psychology*, vol. 25, pp. 210–216, May 1978.

[283] U. M. Marcinkowska, M. V. Kozlov, H. Cai, J. Contreras-Garduño, B. J. Dixson, G. A. Oana, G. Kaminski, N. P. Li, M. T. Lyons, I. E. Onyishi, *et al.*, "Cross-cultural variation in men's preference for sexual dimorphism in women's faces," *Biology letters*, vol. 10, no. 4, p. 20130850, 2014.

[284] P. C. Wason, "On the failure to eliminate hypotheses in a conceptual task," *Quarterly Journal of Experimental Psychology*, vol. 12, pp. 129 – 140, 1960.

[285] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik, "Event registry: learning about world events from news," in *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, (New York, NY, USA), p. 107–110, Association for Computing Machinery, 2014.

[286] A. O. Vladimir Alexiev, Boyan Bechev, "The innograph artificial intelligence taxonomy, white paper." online: https://www.ontotext.com/blog/the-innograph-artificial-intelligence-taxonomy/, 12 2023.

[287] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, pp. 1273–1282, PMLR, 2017.

[288] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1322–1333, 2015.

[289] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks.," in *USENIX Security Symposium*, vol. 267, 2019.

[290] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282, IEEE, 2018.

[291] D. Li and J. Wang, "Fedmd: Heterogenous federated learning via model distillation," in *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*, 2019.

[292] C. Yun-Hin, J. Zhihan, D. Jing, and N. C.-H. Edith, "Fedin: Federated intermediate layers learning for model heterogeneity," *arXiv preprint arXiv:2304.00759*, 2023.

[293] S. Caldas, J. Konečny, H. B. McMahan, and A. Talwalkar, "Expanding the reach of federated learning by reducing client resource requirements," *arXiv preprint arXiv:1812.07210*, 2018.

[294] E. Diao, J. Ding, and V. Tarokh, "Hetero{fl}: Computation and communication efficient federated learning for heterogeneous clients," in *International Conference on Learning Representations*, 2021.

[295] Y. Gu, Y. Bai, and S. Xu, "Cs-mia: Membership inference attack based on prediction confidence series in federated learning," *Journal of Information Security and Applications*, vol. 67, p. 103201, 2022.

[296] D. Bernau, J. Robl, P. W. Grassal, S. Schneider, and F. Kerschbaum, "Comparing local and central differential privacy using membership inference attacks," in *Data and Applications Security and Privacy XXXV: 35th Annual IFIP WG 11.3 Conference, DBSec 2021, Calgary, Canada, July 19–20, 2021, Proceedings*, pp. 22–42, Springer, 2021.

[297] Y. Kaya and T. Dumitras, "When does data augmentation help with membership inference attacks?," in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 5345–5355, PMLR, 18–24 Jul 2021.

[298] M. Shin, C. Hwang, J. Kim, J. Park, M. Bennis, and S.-L. Kim, "Xor mixup: Privacy-preserving data augmentation for one-shot federated learning," *arXiv preprint arXiv:2006.05148*, 2020.

[299] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, "Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays," *PLOS Genetics*, vol. 4, pp. 1–9, 08 2008.

[300] A. Oprea and A. Vassilev, "Adversarial machine learning: A taxonomy and terminology of attacks and mitigations," tech. rep., National Institute of Standards and Technology US Department of Commerce, 2023.

[301] B. Jayaraman, L. Wang, K. Knipmeyer, Q. Gu, and D. Evans, "Revisiting membership inference under realistic assumptions," *Proceedings on Privacy Enhancing Technologies*, vol. 2021, no. 2, 2021.

[302] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18, IEEE, 2017.

[303] L. Song, R. Shokri, and P. Mittal, "Membership inference attacks against adversarially robust deep learning models," in *2019 IEEE Security and Privacy Workshops (SPW)*, pp. 50–56, IEEE, 2019.

[304] G. D. Németh, M. A. Lozano, N. Quadrianto, and N. M. Oliver, "A snapshot of the frontiers of client selection in federated learning," *Transactions on Machine Learning Research*, 2022.

[305] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, IEEE, 2022.

[306] Y. Liu, Z. Zhao, M. Backes, and Y. Zhang, "Membership inference attacks by exploiting loss trajectory," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2085–2098, 2022.

[307] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[308] Z. Li, L. Wang, G. Chen, Z. Zhang, M. Shafiq, and Z. Gu, "E2egi: End-to-end gradient inversion in federated learning," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[309] C. Zhang, Z. Xiaoman, E. Sotthiwat, Y. Xu, P. Liu, L. Zhen, and Y. Liu, "Generative gradient inversion via over-parameterized networks in federated learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5126–5135, 2023.

[310] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership inference attacks on machine learning: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 11s, pp. 1–37, 2022.

[311] G. D. Németh, M. Á. Lozano, N. Quadrianto, and N. Oliver, "Addressing membership inference attack in federated learning with model compression," *arXiv preprint arXiv:2311.17750*, 2023.