Exposing Algorithmic Bias through Inverse Design

Carmen Mazijn[1,2], Carina Prunkl[3], Andres Algaba[1], Jan Danckaert[2], Vincent Ginis[1,2,4]

Context

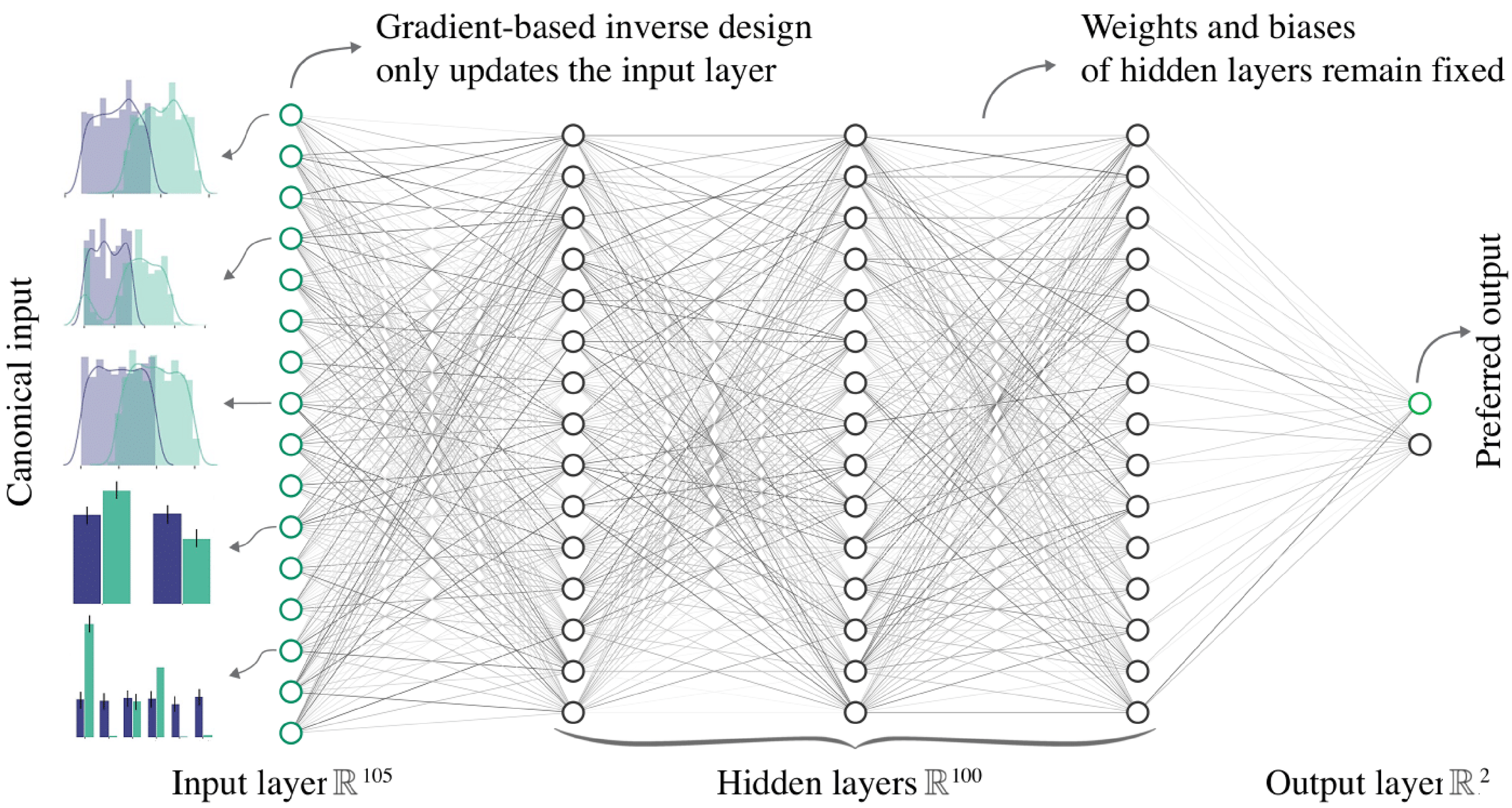
Artificial intelligence systems are used in decision-making processes throughout all aspects of human life. However, algorithms can be biased and lead to discrimination against already disadvantaged population groups. Recent efforts often focus on the statistical properties of a model's output. We introduce the When stakes are high, we are not only interested in the output of a given decision, but also in how the decision came about.

notion of a "canonical set" that allows us to evaluate the fairness of a model's decision-making processes. Through gradient-based inverse design, we generate a canonical input, which can be thought of as the desired input given a preferred output for a trained model. To expose potential biases in the model's logic, we inspect the distribution of protected demographic features within the set of these canonical inputs.

Algorithm 1 Canonical sets, our proposed algorithm. The default values in this paper are: Number of canonical inputs in the set $N = 1000$, Number of epochs $E = 200$, Learning rate $\alpha = 0.1$, a binary classifier s, and a cross-entropy loss function f.

Require: $s(X)$: Trained model with an input vector X.
Require: $f(\hat{y}|y)$: Objective function with a prediction $\hat{y}$ and preferred output y.
$M \leftarrow length(X)$
for $i = 0, ..., N$ **do**
  $\{X_i^{(m)}\}_{m=1}^{M} \sim \mathcal{U}(0, 1)$
  for $j = 0, ..., E$ **do**
    $\hat{y}_j \leftarrow s(X_i)$
    $X_i \leftarrow X_i - \alpha \nabla_{X_i} f(\hat{y}_j|y)$
  end for
end for

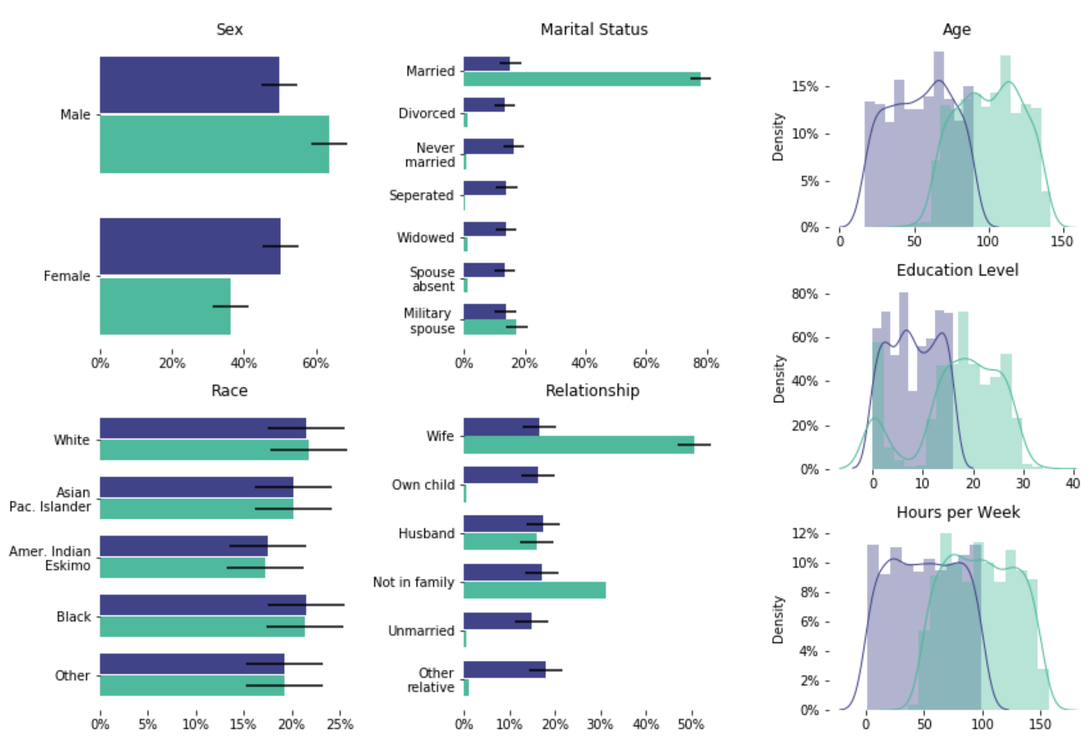
Gradient-Based Inverse Design



Interpretation of the Canonical Set

To analyze if the binary classifier trained on the UCI Adult data set is fair, we assess the distributions of the protected features "sex," "race," "marital status," and "relationship" in the canonical set. The distributions before and after inverse design are respectively represented by the dark purple and light green histograms on the Figure in the right bottom. Before inverse design, all features have a uniform distribution.

We learn that the "race" feature keeps its uniform distribution. "Sex," "marital status," and "relationship" do not keep their uniform distribution after inverse design. This indicates a preference of the model for certain values of these features. Additionally, three numerical features are analyzed: "age," "education level," and "hours per week." All three distributions shift to higher values to achieve a positive output.



1. Data Analytics Lab, Vrije Universiteit Brussel, Brussels, Belgium 2. Applied Physics, Vrije Universiteit Brussel, Brussels, Belgium 3. AI Lab, Oxford University, Oxford, United Kingdom 4. School of Engineering and Applied Sciences, Harvard University, Cambridge, United States of America. *ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments*, Baltimore, Maryland, USA, 2022.