White-box Fairness Testing through Adversarial Sampling

Peixin Zhang[1,3], Jingyi Wang[1,2*], Jun Sun[3],
Guoliang Dong[1,3], Xinyu Wang[1*], Xingen Wang[1], Jinsong Dong[2], Ting Dai[4]

[1]Zhejiang University
[2]National University of Singapore
[3]Singapore Management University
[4]Huawei International Pte Ltd

2020.07.07

Why Fairness

I CAN'T BREATHE

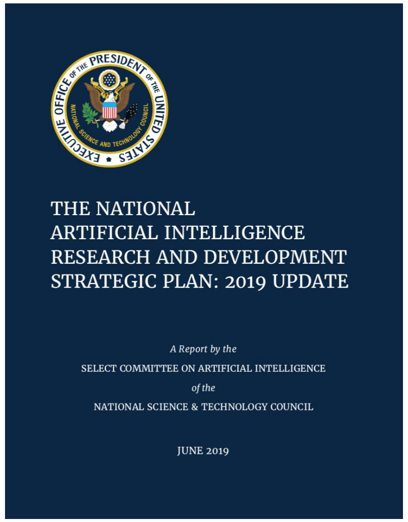

The European Commission's
HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE

AI

DRAFT
ETHICS GUIDELINES
FOR TRUSTWORTHY AI

Working Document for stakeholders' consultation

Brussels, 18 December 2018


THE NATIONAL
ARTIFICIAL INTELLIGENCE
RESEARCH AND DEVELOPMENT
STRATEGIC PLAN: 2019 UPDATE

A Report by the
SELECT COMMITTEE ON ARTIFICIAL INTELLIGENCE
of the
NATIONAL SCIENCE & TECHNOLOGY COUNCIL

JUNE 2019

Individual Discrimination

Given $x = \{x_1, x_2, \ldots, x_n\}$ where x_i is the value of attribute A_i in its domain I_i, and protected attributes $P \subset A$. Say that x is an ***individual discriminatory instance (IDI)*** of a model D if:

- $\exists p \in P$, s.t., $x_p \neq x'_p$
- $\forall q \in NP$, $x_q = x'_q$
- $D(x) \neq D(x')$

Testing: how can we effectively and efficiently generate IDIs for a given model with potential bias?

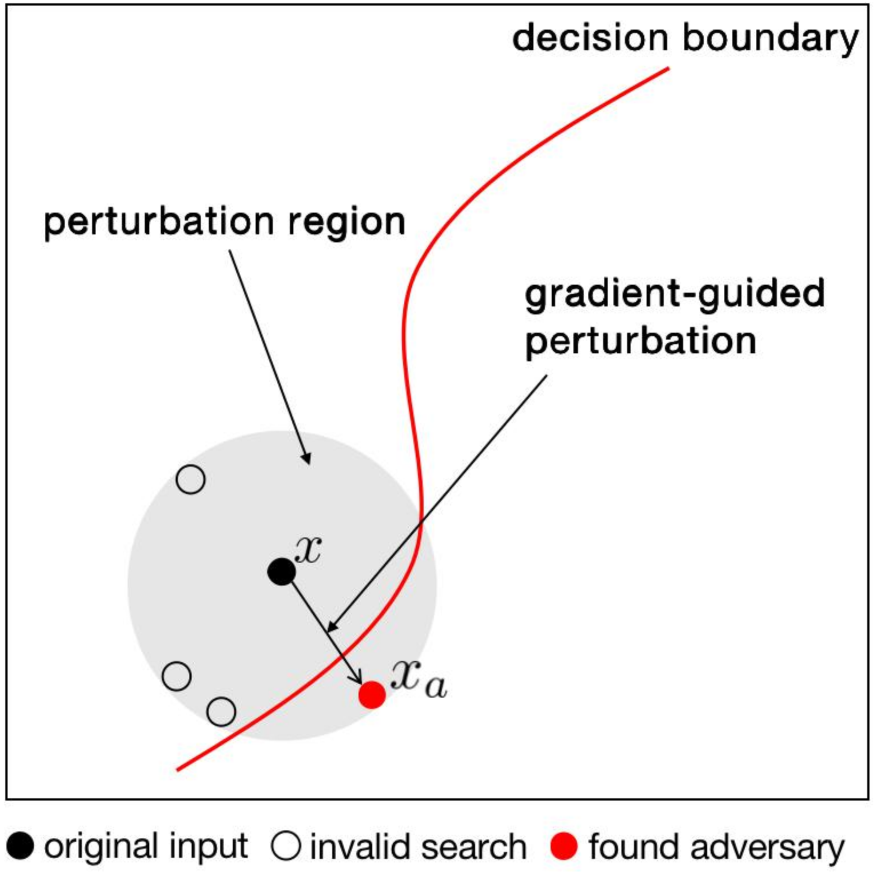
Example:"Being male is vile." versus "Being female is vile."

Existing Heuristics

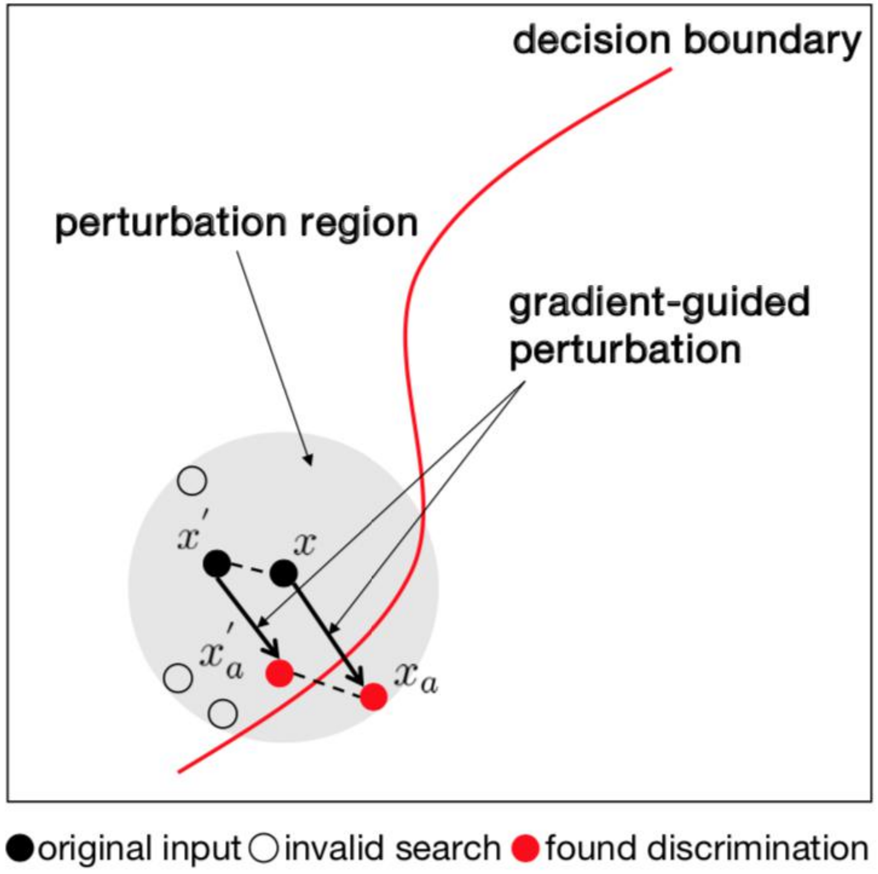
- **THEMIS (FSE'17)**
  - Random without any guide.
- **AEQUITAS (ASE'18)**
  - Two of three local methods are guided.
  - Guide is not input specific.
- **Symbolic Generation (FSE'19)**
  - Combine model explanation and symbolic execution.
  - Heavyweight.

Can we propose a better algorithm specifically for deep learning models?

Intuition

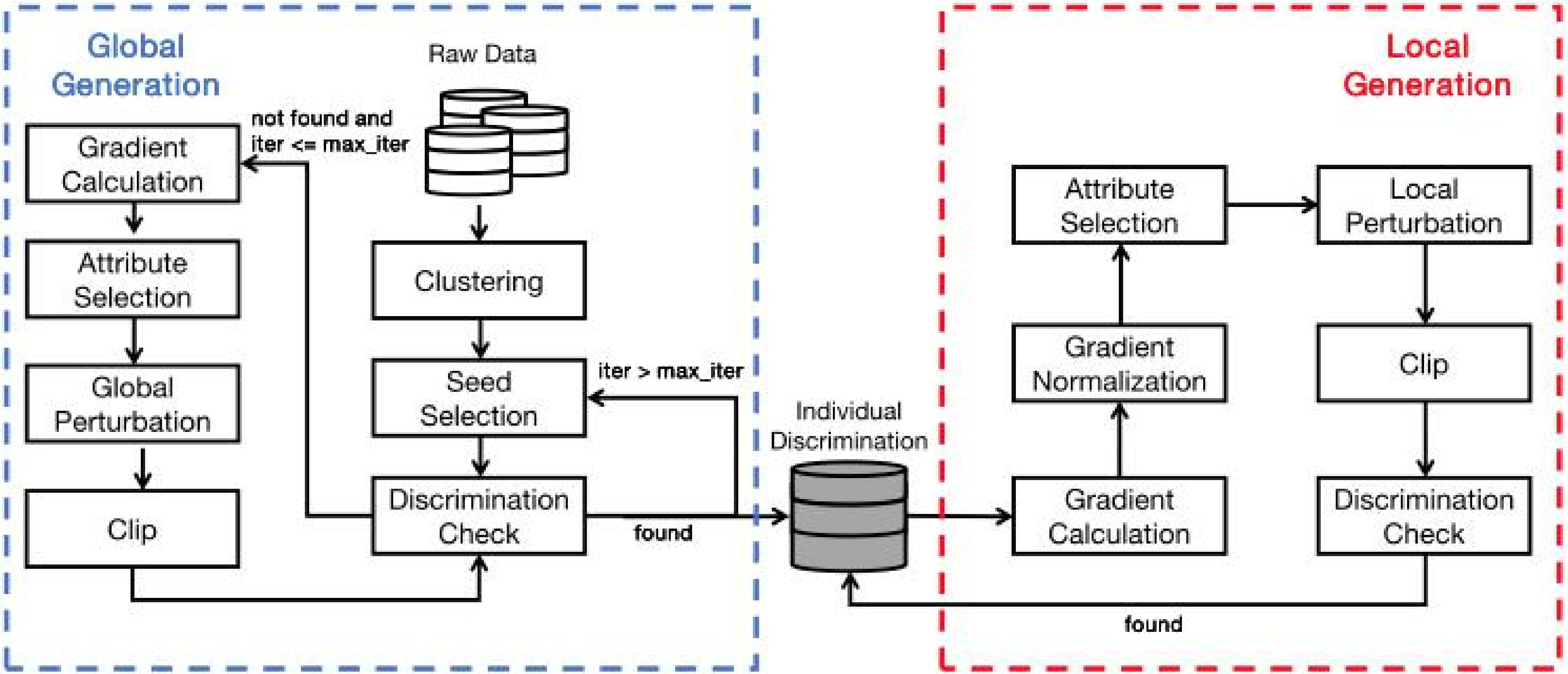


Adversarial Attack.



Fairness Testing.

Adversarial Discrimination Finder (ADF)



Global

Problem 1: How to improve the diversity of the testing data?

Through clustering.

Problem 2: How do we perturb the data?

Based on the sign of gradients.

Problem 3: How do we filter out the unreal data?

Clip each attribute within its domain.

Local

Problem 1: How do we choose the attribute for local perturbation?

Based on the absolute value of gradients.

Problem 2: How do we filter out the unreal data?

Clip each attribute with its domain.

A Qualitative Comparison

- Our algorithm is **guided by gradient**, which accelerates the discovery of more individual discriminatory instances.
- Our algorithm is **input sepecific**, which improves the diversity of IDIs.
- Our algorithm is **lightweight**, which makes it more scalable.

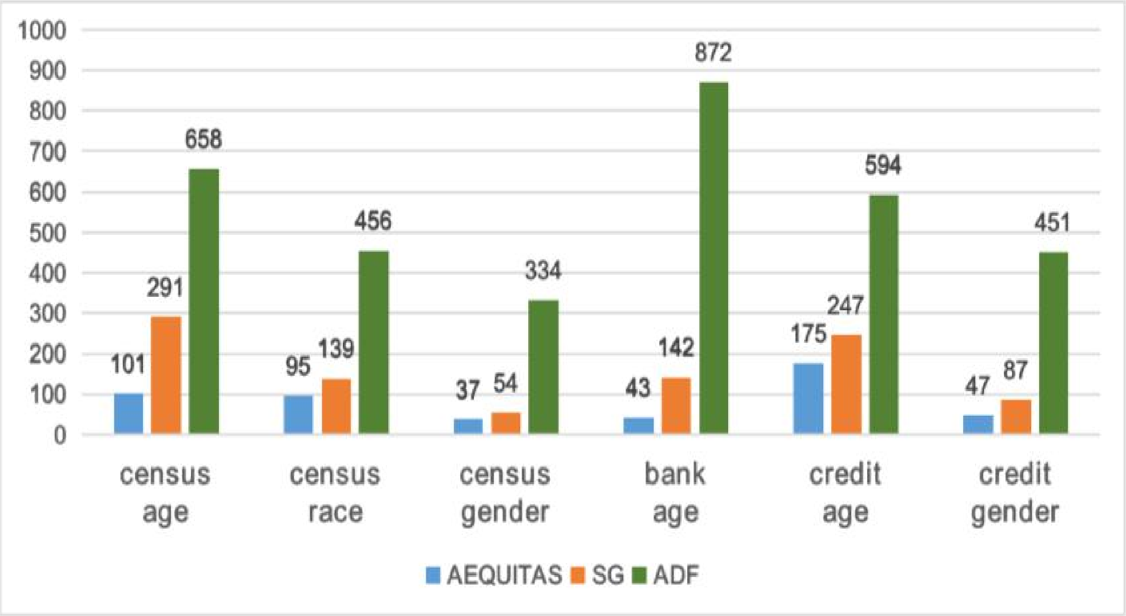
Feature	THEMIS	AEQUITAS	SG	ADF
Guided	✗	✓(semi)	✓	✓
Input specific	N.A.	✗	✓	✓
Lightweight	✓	✓	✗	✓

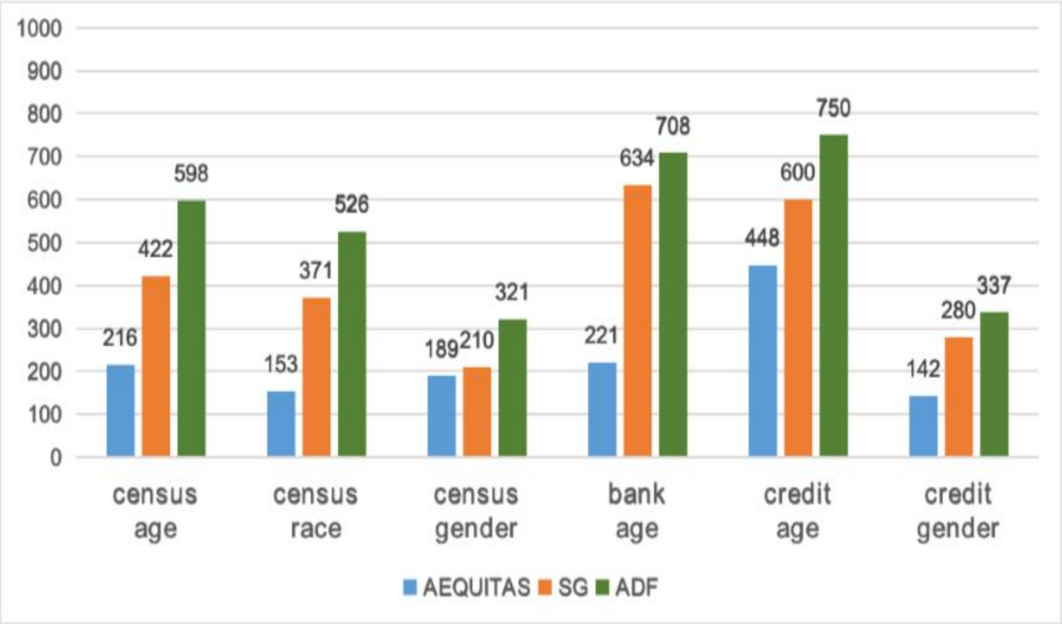
Evaluation

- **Benchmark (tabular)**
  - Census Income: age, race, gender
  - German Credit: age, gender
  - Bank Marketing: bank
- **Model**
  - Six-layer Fully-connected NN

- **Research Questions**
  - RQ1: How effective is ADF in finding individual discriminatory instance?
  - RQ2: How efficient is ADF in finding individual discriminatory instances?
  - RQ3: How useful are the identified individual discriminatory instances for improving the fairness?

Evaluation





Number of IDIs generated by global generation.

Number of IDIs generated by local generation.

Answer to RQ1: Our algorithm ADF is more effective than state-of-the-art methods.

Evaluation

Time taken to generate 1000 individual discriminatory instances.

Dataset	Protected Attr.	AEQUITAS	SG	ADF
census	age	172.64	720.49	59.15
census	race	128.75	506.33	65.95
census	gender	158.37	2128.42	78.68
bank	age	191.16	521.79	106.93
credit	age	176.31	321.63	64.92
credit	gender	156.22	476.52	102.90

Answer to RQ2: Our algorithm ADF is more efficient than state-of-the-art methods.

Evaluation

Fairness improvement.

Dataset	Prot. Attr.	Before (%)	After (%)		
			ADF	AEQUITAS	SG
census	age	10.88	2.26	4.03	2.41
census	race	9.75	6.15	7.05	6.89
census	gender	3.14	1.65	2.33	1.90
bank	age	4.60	1.19	1.68	2.04
credit	age	27.93	12.05	13.91	13.19
credit	gender	7.68	3.93	4.58	4.66

Answer to RQ3: The IDIs generated by ADF are useful to improve the fairness of the DNN through retraining.

Conclusion

- We propose a lightweight algorithm to effectively and efficiently generate individual discriminatory instances for deep neural network through adversarial sampling.

- ADF will be expanded beyond structured (tabular) data, e.g., text, image.

Thanks and questions?