Dataset Generation

The CSV files and python scripts to generate the datasets for each study are included.

Study 1: *dataset_formative1.csv* and *data_generator_formative1.py*

Study 2: *dataset_formative2.csv* and *data_generator_formative2.py*

Main Study: *dataset_main_study.csv* and *data_generator_main_study.py*

Each script takes a number *n* representing the number of rows (politicians) to generate. The script then generates one politician at a time, appending it to a list, then writing the complete list to a file line-by-line.

An individual politician is generated by sampling each attribute value from the distributions defined in the script. The distributions for each attribute vary per study and can be found in the respective scripts. For Study 1 and the Main Study, we sought for the dataset as a whole to adhere to specific controlled distributions -- hence, for these studies, we generated additional rows (politicians) for underrepresented attributes and pruned rows for overrepresented attributes until the desired distribution was achieved.