## Youtube Sentiment Analysis on AI in Europe

Our group proposes an experimental project that involves conducting sentiment analysis on YouTube videos about AI using their comments.

**Description of the Task:** Sentiment Analysis is a technique used to determine the sentiment expressed in a piece of text. In the context of our project, the goal is to analyze the comments on YouTube videos about AI and determine the sentiment towards AI. This process will be repeated for videos from the largest countries in Europe. The results from each country will then be compared to understand the differences and similarities in sentiment towards AI across Europe. To achieve these results, we will use a pre-trained model and fine-tune it for our specific task. The analysis should provide insights into the public opinion about AI in these countries, and the ambition is that this project could be valuable for AI researchers, policymakers, and businesses operating in the AI sector in Europe.

# Obtaining Dataset

First thing first we need a dataset to fine tune the chosen model. We will obtain our data from YouTube comments.

1. Build several excel files to store the informations of the videos to scrape. (Daniele Virzì)
   1. Number of comments.
   2. Title.
   3. Link.
2. Identify search queries to use on YouTube, dividing them in “Positive”, “Neutral” and “Negative” Bias on AI based on the title and the content of the videos. (Everyone)

POSITIVE: pros of ai, why ai won’t replace humans

NEUTRAL: ai and its impact on society, ai will change the world, future of ai, ai and its ethics, ai and its implications, artificial intelligence

NEGATIVE: cons of ai, why ai stole my job, ai is a threat, why ai will replace humans.

1. Search several videos with these inputs, aiming to get more than 100k comments in English (Training and Test) and 10k for the countries (Inference). (Everyone)
2. Implemented a YouTube comment scraper using YouTube v3 APIs. (Daniele Virzì)
3. Applied pre-processing to remove emojis, NAs, stop-words, normalise font if special characters are present, and turn it in lowercase. After that we applied lemmatising and used buzz words to identify the important comments. (Marlon Helbing, Daniele Virzì)

# Model

We then need a model trained on sentiment analysis to fine tune for our specific task.

1. Look for an already trained model to fine tune, in our case distilBERT – base – multilingual – cased – sentiment. (Daniele Virzì)
2. Use this model to label comments and only take the ones classified with x% accuracy threshold. (Marlon Helbing)
3. Manually check the labels and if needed modify them, to introduce our bias. (Everyone)
4. Translate these labeled comments to fine tune a specific model on ai sentiment analysis for each language. (Marlon Helbing)
5. Finally use the models to inference different languages. (Everyone)

Before deciding for this approach, we tried manually labeling 10k English comments and use them to fine tune the model on ai sentiment analysis and do step 4 but we realized not only the model’s predicted labels were different from ours, but also that, as we saw in class, we often had a different label opinion for the same comment.

# Inference

Now we can start our inference on different EU countries and obtain their perceived image of ai.