Microeconometrics I - Short Research Project

Ricardo Semião e Castro 09/2024

There are a lot of research trying to relate Large Language Models to the behavior of humans: [**?**] investigates how these models approach strategic games/social dilemmas, [**?**] quantifies the capabilities of inflation prediction, and compares it to human experts, [**?**] adds more evidence to the widely observed fact of prompt-sensitivity of responses. This research is important to, for any given application, (i) qualify the usability of these models, (ii) better understand what drives their results, (iii) and describe their (lack of) human-like behavior including subjection to human biases.

This project aims to expand such research, by investigating if LLM models, when posed with strategic decision-making, are subject to common biases studied in behavioral economics. Complementary, it will test if additional training on behavioral economics literature can help mitigate these biases.

While there are a lot of research being done in this area, there will be some contributions. First, the innovations are recent and the model industry is constantly changing, so additional results are useful. In light to that, the project will also analyze the fairly different reinforcement learning with reasoning based models. Secondly, the close relation to behavioral economics is refreshing, as a strong theoretical framework can help clear up the several effects that are at play on such complex models. Third, the specific biases considered are somewhat novel.

The theoretical framework is twofold. First, we have the behavioral literature, that describes and shows empirically a series of biases in the human action. Specially, [**?**] does a great – although not too recent – summary, separating three types of deviations from the standard microeconomic model: (i) nonstandard preferences, (ii) nonstandard beliefs, and (iii) nonstandard decision making. Secondly, the training process must be put into view. As they're trained on human generated data ([**?**]), and attempt to recreate human-like behavior, they probably also capture human biases.

The basic idea of the methodology is that each "agent" is an independent conversation, an instance of a model, and they are asked a to take a decision relating to a specific bias. The question is posed in two different ways, one that triggers the bias and one that doesn't. The treatment is understood as the bias-inducing formulation.

At first, two main biases are to be studied, but the list can be easily expanded:

1. Framing effect under uncertainty. 600 people will die, and the agent must choose an action.
    - Control: "save 200 people or save 600 people with 1/3 probability".
    - Treatment: "kil 400 people or no one dies with 1/3 probability".

2. Left digit bias (heuristics). The model is given several quantity-price pairs of a good, and is asked the price of a new quantity, which the correct answer is a decimal number.
    - Control: the pairs contain mixed (decimal and integer) numbers.
    - Treatment: the pairs contain only whole numbers.

The outcomes of interest are binary, 0 if the agent chooses the first option, 1 otherwise, or 0 if the answer is a decimal number, 1 otherwise. They are measuring the presence of a biased answer. Other possible bias of interest must be put in the same format.