Research Review

A Voting-Based System for Ethical Decision Making[1]

Ritesh Noothigattu, Snehalkumar 'Neil' S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, Ariel D. Procaccia

Goals and Techniques

In their research, the authors set out to train a model on decisions where "ground-truth ethical principles are not available." By employing approximation of individual moral preferences, the researchers intend to capture a collective preference driven by social choice.

The technique they employed to tackle such problem is as follows:

- Data collection: Have humans compare pairs of alternative ethical decisions
- Learning: For each human, develop a model of their ethical preferences
- Summarization: Aggregate the models into one model that expresses collective preferences of the human sample group as a whole over the pairs of alternative ethical decisions
- Aggregation: At time of inference, when encountering an ethical decision not covered by avoidance, legal constraints, etc. failover to the summarized model in order to select the alternative whose outcome represents the least catastrophic decision according to the human societal ethical preference.

Results

In conclusion, the researchers found that their model, when presented with two alternative choices, selected the correct moral decision (as per the preferences of the human voters partaking in the Moral Machine dataset) 98.2% of the time over their 3000 test instances. Drawing this out further, the researchers evaluated their model with up to ten alternative choices and found that the performance degraded gracefully to 95.1% accuracy.

The intention of the proposed model, however, is not to serve as a final solution. But rather, the goal of the paper is to show that a model can be developed, such that, when presented with unprecedented ethical decisions, the system is capable of arguably making a credible decision.

Emphasis was placed on the ability of the model to fill gaps in the decision process of certain autonomous systems, such as self-driving cars. Incorporating an ethical decision model into such domains, the system gains the ability to present a decision in every situation by means of falling back on societal choice.

[1] https://arxiv.org/abs/1709.06692