IST 402
Week 11 Notes
Alex Kim, Ben, Rommel

AI Based Ethical Frameworks

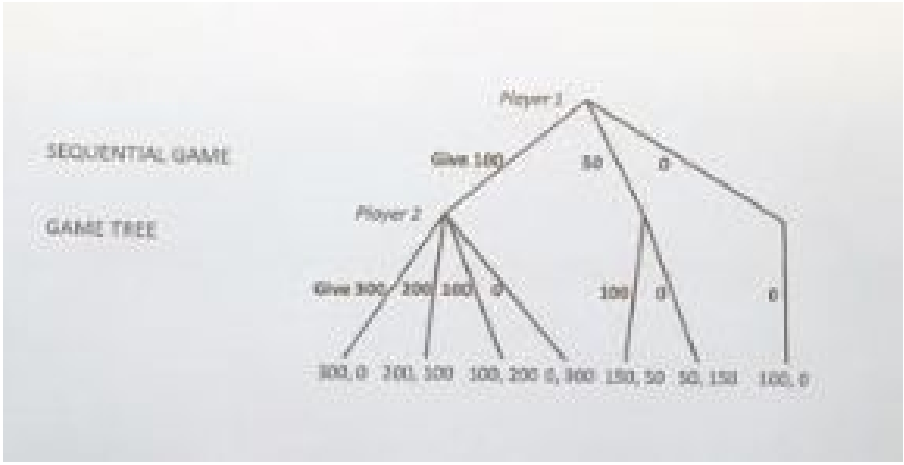
- 2 Different ways to frame moral AI
    - Artificial Intelligence Models
    - Game Theory
    - A way to combine the two
- Game Theoretic solvers
    - Trust game
    - two players
    - Player 1 is given some amount of money -- say $100
    - She is then allowed to give any fraction of this money back to the experimenter
    - Experimenter triples this returned money and give it to player 2
    - Finally, player 2 may return any fraction of the money he has received to player 1
        - What is player 2's optimal choice?
            - Give no money
            - Player 1 can't do anything
        - What is players 1's optimal choice
            - Give no money
            - Player 2 can't do anything
            - Both decisions are rational perspectives

- How to play trust game?
    - Each player, at any point in the game is interested only in maximizing the amount of money she herself receives
    - Under this assumption, player 2 would never have reason to return money to player 1
    - Anticipating this, player 1 would not give any money either

- What existing work in trust game shows
    - Human subjects playing the trust game generally do give big money in both roles
        - Why?
    - Many people feel it is wrong for player 2 not to give any money back after player 1 has decided to give him some (and, when in the role of player 1, they expect player 2 not to take such a wrong action).

- What existing work in trust game shows?
    - Most people consider not only the consequences of their actions but also the setting in which they perform their actions
    - They ask whether an act would be

        - Unfair or selfish (because they are not sharing a good with someone who is equally deserving)
        - **Ungrateful (because it harms someone who benefited them in the past)**
        - Disloyal (by betraying someone who has been loyal)
        - Untrustworthy (because it breaks a promise)
        - Deserved (because the person won a competition or committed a crime)
    - In these ways, moral reasoners typically look not only to the future but also the past
    - Standard Game theory does not account for any of these considerations

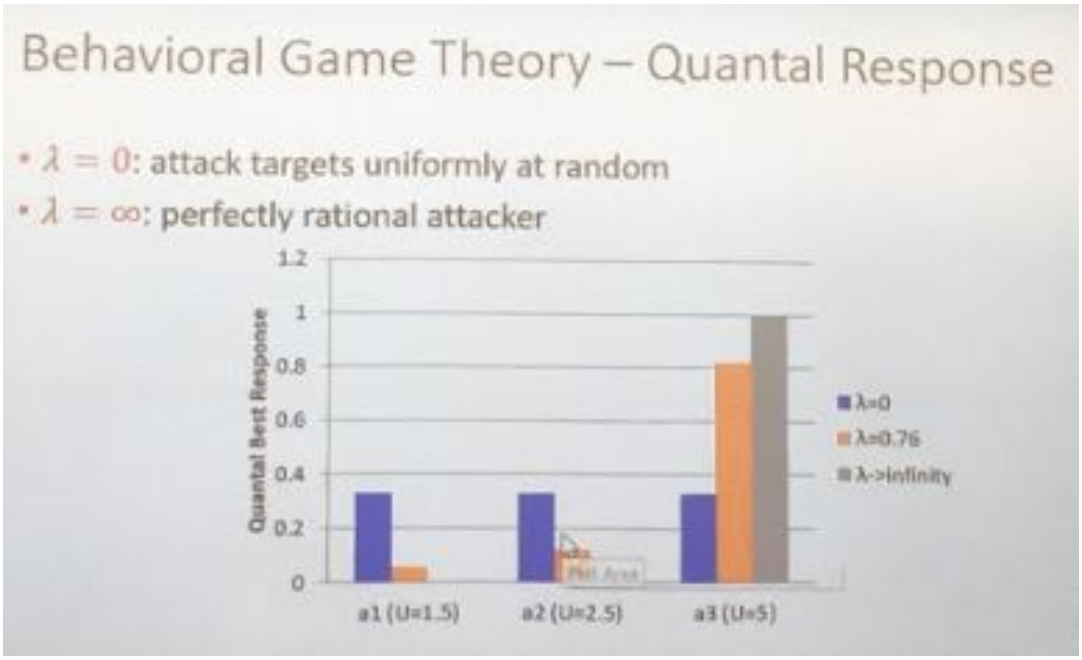
- How to model the trust game in game theory
    - W/ a sequential game - game tree
    - Depending on the choices made, there will be a node created leading to different decision nodes and outcomes
    - 

    - Backward induction is standard method to find optimal decision

- However this is not the behavior observed in experiments with human subjects
- Many games that elicit human behaviors do not match game theoretic analysis such as the trust game
- Often used to criticize the game theoretic model of behavior
- Led to the field of behavioral game theory
    - Models boundedly rational human beings as opposed to perfectly rational players, which is the standard assumption in game theory

- Behavior game theory - quantal response
    - Humans are not infinitely rational and cannot be expected to perform complete game-theoretic analysis in their heads
    - Error in individuals response
        - Still: more likely to select better choices than worse choices

    - Probability distribution of different responses
    - Boundedly rational attacker: attack a target ti with probability:

- Behavioral Game Theory - Quantal Response
    - 

    - Modelling the fact that humans find it difficult to compute the best actions
        - They make mistakes, so as a result of which they might pick sub-optimal actions with a small probability

- Behavior Game Theory Does not help...
    - Not the primary reason that agents behave differently in the trust game
    - If they don't give any money back, they will betray player 1, who voluntarily decided to give money back to the experimenter
    - Simplistic game -theoretic solutions fails to account for ethical considerations

- Can you change the reward functions of the players?
    - An agent’s utility may take into account the welfare of others, so it is possible for altruism to be captured by a game theoretic model
    - What is morally right or wrong also seems to depend on pat actions by other players
        - E.g. betrayal
        - If another agent knowingly enables me either to act to benefit us both, or to act to benefit myself even more while significantly hurting the other agent, doing the latter seems morally wrong
        - What does this mean for the trust game?

        - This cannot be modeled by existing game theoretic methods

- One Possible Solution - Conitzer et al.
    - This solution concept involves repeatedly solving the game and then modifying the agents' preferences based on the solution
    - **Key Insights**: Player 2 wants to ensure that player 1 receives at least what she could have received in the previous solution, unless this conflicts with player2 receiving at least as much as he would have received in the previous solution
        - Workable definition of morality that can be embedded inside AI algorithms
            - benefit others as much as possible without hurting yourself
    - In the trust game player 2's preferences are modified so that he values player 1 receiving back at least what she gave to player 2 without hurting his own preferences

- Only the tip of the iceberg..
    - This concept has been defined for only restricted settings
        - 2 player perfect information games - Trust game
    - Future directions are necessary to make these solutions more useful in the real-world
    - Need to generalize the concept to games with more players and imperfect information
    - Need to define different solution concepts that capture other ethical concerns.

- Limitations of the Game Theory Approach
    - Game theory can capture much of what is relevant in ethics, but they may not capture everything that is relevant
    - However, many philosophers, would argue that there is a significant distinction between the two alternatives, and that switching train to the second track is morally wrong
    - For example: trolley problem
        - if the trolley goes straight, it will kill a lot of people
        - if the lever is pulled, it will go on another track and kill only 1 person
    - X
    - X
- The machine learning approach
    - Another approach for developing procedures that automatically make moral decisions is based on machine learning
    - We can assemble a training set of moral decision problem instances labeled with human judgements of the morally correct decisions, allow our AI system to generalize
        - Moral machine experiment is 1 example of such a dataset
    - How is each situation represented?
        - It is insufficient to represent the instances in natural language

        - We must represent them more abstractly

- Example: Moral Machine Experiment
    - In the moral machine experiment, what could be relevant features?
        - whether people were following rules of not
        - Amount of deaths in both situations
        - whether human beings or pets
        - The particular condition in which the human beings are (e.g. pregnant lady as opposed to lady)
        - The output that we are getting from the game-theoretic framework (binary output, decision is ethical or not) or (probabilistic output)
    - The primary goal of a general framework for moral decision making is to identify abstract features that apply across domains, rather than to identify every nuanced feature that is potentially relevant to isolated scenarios

- Key future directions
    - Given a labeled dataset of moral dilemmas represented as lists of feature values, we can apply standard machine learning techniques to learn to classify actions as morally right or wrong.
    - Often seen as important not only to act in accordance with moral principles but also to be able to explain why one's actions are morally right
        - Interpretability
    - X
    - X
- Key Future Directions
    - Is decision making binary in this domain? Is one answer always strictly more moral than another answer
    - Why should the labels be by binary then?
        - Why can't the labels reflect our uncertainty about
    - Idea 1: we may make a quantitative assessment of how morally wrong the action is using regression
    - Idea 2: we may make an assessment of how likely it is that the action is morally wrong using Bayesian frameworks

- How to combine the two approaches
    - We can apply moral game -theoretic concepts to moral dilemmas
    - Use the output (say right or wrong according to this concept) as one of the features in our machine learning approach
    - The outcomes of the machine learning approach can help us see which key moral aspects are missing from our moral game theoretic concepts, which will in turn allow us to refine them

- Can ML approaches do better than human moral ethics?

- ML approaches to moral decisions will be limited because they will at best result in huan-level moral decision making; they will never exceed the morality of humans
- Why is that a bad thing?
  - It has not been established that the human notions of morality are bad, and hence mimicking them is not necessarily a bad decision.
- Aggregating the moral views of multiple humans in ML may result in a morally better system than that of any individual human
- Idiosyncratic moral mistakes made by individual humans are washed out in the aggregate
- ML approaches may identify general principles of moral decision making that humans were not aware of before

A COMPUTATIONAL MODEL OF COMMONSENSE MORAL DECISION MAKING