Tuesday 10/29/19

The Moral Machine Experiment

Iyad Rahwan's Ted Talk

· Talk centered around technology and society, and self-driving cars - social aspects of technology

· Moral decision making

· Bentham line of thought: minimize death count

· Dilemma: do you take action to minimize the death count? Or do not take action and let the situation play out

· Social dilemma: if everyone thinks, "I won't follow the rules, but I hope everyone else will", then no one will follow the rules

  - This is why collecting user data on ethical preferences is important

· Asimov's laws of robotics

  - A robot may not injure a human being or allow a human being to come to harm.
  - A robot must obey the orders given to it by human beings except if they were to conflict with the first law.
  - A robot must protect its own existence - may not allow itself to come to harm

Moral Machine Experiment

· Collected data based off culture clusters

  - Western, eastern, southern
  - Eastern cares more about people following the law, less about sparing the young, don't follow Bentham model
  - Should be taken into account by regulators so the laws that are formed are modeled based on the preferences of that country

· Collected data from various regions and made charts to show probabilities of choosing one side over another in a situation where an ethical decision is made, in context of autonomous vehicle accidents

Moral Machine Experiment (continued)

· Study, gathered 40 million decisions in 10 different languages from many people spanning across 233 countries

  - Moral preferences among those who participated in the study: prioritize young lives, valuing humans over animals
  - Spared most often were babies in strollers, pregnant women, and children
  - Final results also showed ethics varying between different cultures

• Let the car take its course. Even if it will harm more people. Do not take an action that will harm others.

• If the man pulls the lever, he makes the conscious choice and kills a person. If he doesn't pull the lever, he had nothing to do with the trains original path.
• Human beings are not rational. We are controlled by emotions.
• People themselves would want a car that protects themselves but wants everyone else to buy a car that will protect others.

Thursday 10/31/19

Prisoners dilemma:
Each player's choice: Cooperate or Defect
If both cooperate, one chocolate each
If both defect, no chocolates for anyone
If one defect, one cooperate: Defect gets two chocolates, cooperate owes me a chocolate

Defect is the safest option. You either get nothing or gain 2. If you cooperate you either get 1 or lose 1.

Game Theory for Security Overview

- A means in which any interaction between any 2 agents where each agent has its own interests in mind when making decisions

Prisoner's Dilemma History

- A situation in which two players both have two options, whose outcome depends on the choices made by the other
- Rational decision is to always defect for your best interest, in that the risk reward is greatest for this decision where you either go free or you get 1 year

• Lost Angeles
• Originally created by Merrill Flood and Melvin Dresher while working at (RAND)
    ○ RAND is an American nonprofit created in 1948
• Models many phenomena
• Nuclear arms rivalry:
    - US vs USSR (1950s)
    - Build H-bomb or not
    - Each nation prefers: build h-bomb and the other does not
    - Little gained if both build h-bomb (money wasted)
    - Game theory predicts: both build h-bomb

- John von neumann - Hungarian-American mathematician and computer scientist
    - Normal form games
        - List of players, strategies, payoffs
        - Simultaneous
        - Zero-sum
        - Rock paper scissors
    - Nash equilibrium

- Set of strategies for each player within a game, such that no player has any reason to change his or her strategy based on how the other players are playing
- Can show that decision makers follow randomized strategies