

# Logistics: Project

- Project presentation
  - Dec 7/9 during lectures
  - Everyone is expected to attend both lectures
  - 10 minutes for presentation + 1~2 minutes for QA and transition
- Project reports
  - Due: Dec 12 (no late submissions)
  - Up to 6 pages (plus additional pages for only references/citations)
  - No strict format requirements
    - You are encouraged to use standard templates
- Check Piazza posts for details/updates

# Assignment 4

- No class on Nov 23, Tuesday (happy thanksgiving!)
- Instead, I'll give a list of talks/tutorials relevant to this course.
  - Choose one of them, watch the content, and write a report
  - The report needs to be no less than 2 pages, with any reasonable format
  - The report will serve as your assignment 4
- More details will be announced later this week

# Peer Review

- Please submit the peer review [by 6pm](#)
- Use the link on the schedule of last Thursday

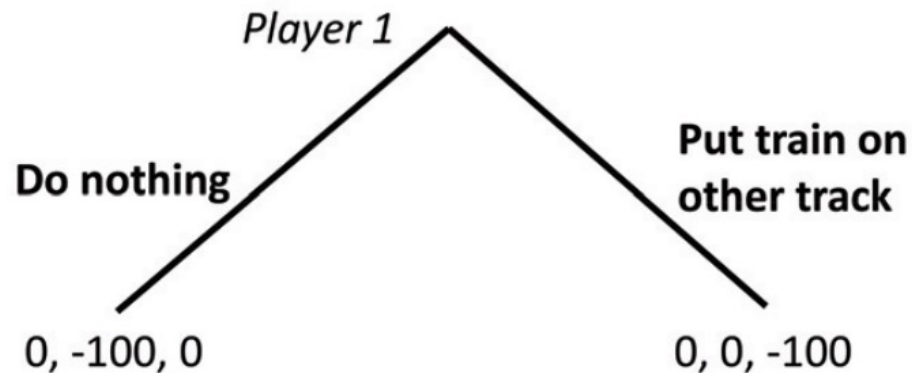
## Lecture 20

Continue on Ethical Decision Making  
Computational Social Choice

Instructor: Chien-Ju (CJ) Ho

# Algorithmic Ethical Decision Making

- A Top-Down Approach
  - Extending game theory / decision theory to incorporate ethics
    - Define what you mean by being ethical
    - Wire that in into the algorithm design
- If you can write down the “moral” utility...



# Asimov's Three Laws of Robotics



- First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

# Algorithmic Ethical Decision Making

- Trust Game
  - Two players: Alice and Bob
  - Game flow
    - Alice receives 100 dollars and can decide to give back  $X$  dollars
    - Bob will then receive  $3 * X$  dollars and can decide to give Alice  $Y$  dollars
- What's the equilibrium?
  - $X = Y = 0$
- In practice, people do give back money [Berg, Dickhaut, and McCabe 1995]
  - It feels morally *wrong* to not give some money back
  - If we can quantify this sense of moral, we can incorporate that in the utility
  - It's usually really hard to do so and everyone is different

# A Related Concept

- A utilizing maximizing algorithm that respect given moral principles

maximizing *utility(action)*  
subject to *moral principles*

- It can often be equivalently represented as follows

[Using Lagrangian idea in convex optimization]

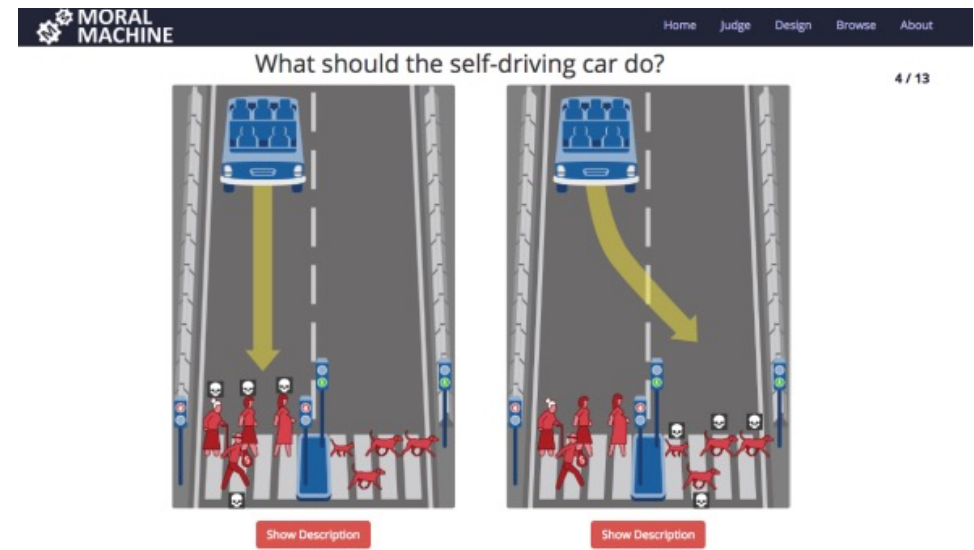
maximize *utility(action) -  $\lambda$  \* [moral principles]*

- Informally, these ethical/fairness discussions often lead to defining utility



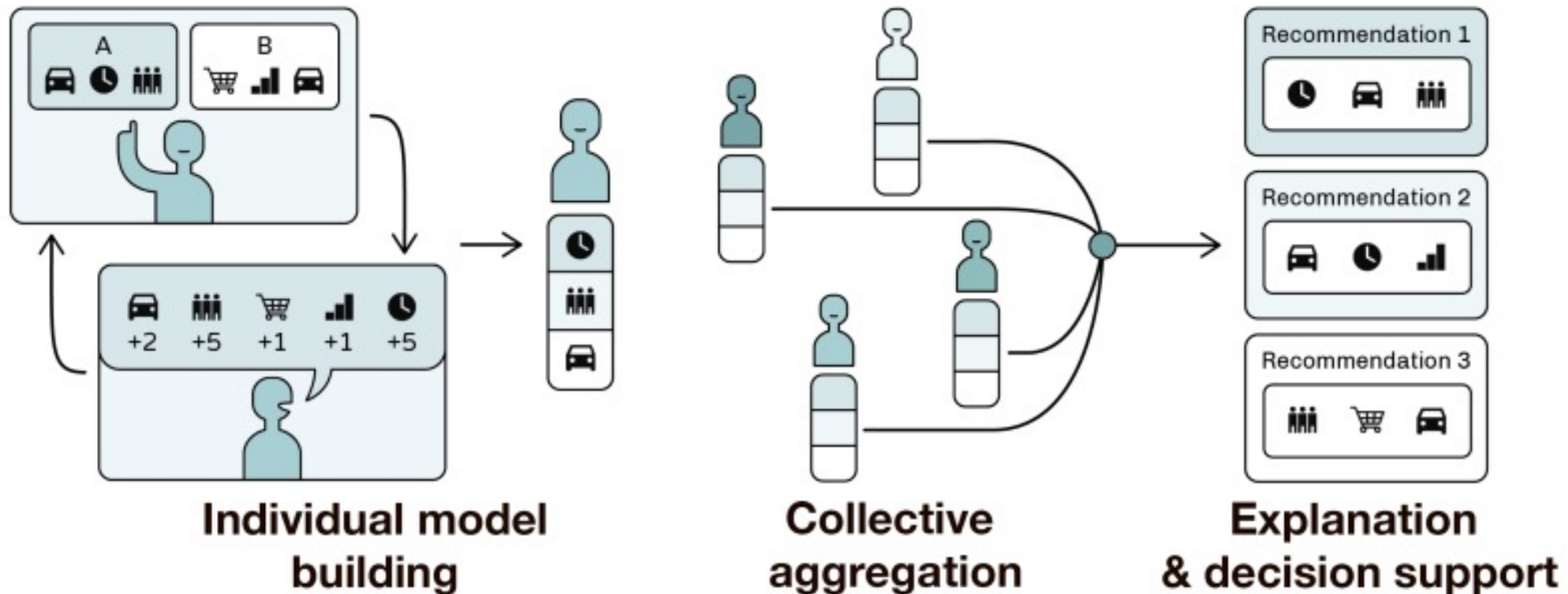
# Algorithmic Ethical Decision Making

- A Bottom-Up Approach
- Learning/summarizing from human judgements
  - Collect data sets of human judgments
  - Apply machine learning
  - Make decision based on predictions



# Incorporating Preferences from Stakeholders

- WeBuildAI: A Participatory Approach [Lee et al. 2019]



# What if People Disagree?

- Social choice theory
  - Given a set of preferences on actions, what is the action we should take?

$\triangle \succ_1 \bigcirc \succ_1 \square$

$\square \succ_2 \triangle \succ_2 \bigcirc$

$\bigcirc \succ_3 \square \succ_3 \triangle$

---

?

# Voting – One Common Form of Social Choice

- Consider the following voting rules (there are a lot more...)
  - **Plurality**: elect the candidate ranked first most often
  - **Borda**: assume  $m$  candidates, each voter gives  $m-1$  points to the candidate she ranks first,  $m-2$  to the candidate she ranks second, etc., and the candidate with the most points wins
  - **Approval**: voters can approve of as many candidates as they wish, and the candidate with the most approvals wins
- Consider this example (bold blue indicates “approval”)

5 voters think: **A**  $\succ$  B  $\succ$  C

4 voters think: **C**  $\succ$  B  $\succ$  A

2 voters think: **B**  $\succ$  **C**  $\succ$  A

Who wins the election for each of the voting rule?

We need to be very clear about what properties we are looking for

# Arrow's Impossibility Theorem

- Some nice criteria we might want to have:
  - **Unanimity**: If every prefers A over B, then the group prefers A over B
  - **Non-dictatorship**: no person's preference is always strictly preferred than others
  - **Independence**: If for two sets of preferences, A and B have the same order between sets, A and B should have the same order in the group decision
- Arrow's Impossibility Theorem
  - No mechanism satisfies the three criteria when the number of candidates  $\geq 3$

# Things could be even more complex

- Consider the artificial scenario, assume we use plurality voting

40% voters think: **Trump** > **Biden** > **Sanders**

35% voters think: **Biden** > **Sanders** > **Trump**

25% voters think: **Sanders** > **Biden** > **Trump**

- If everyone vote according to their preferences, **Trump** wins.
- Supporters of **Sanders** might strategically change their votes to **Biden** to prevent **Trump** from winning

# Axiom Based Approach

- Define a set of axioms that we like
  - **Unanimity, Non-dictatorship, Independence, Strategy-proof,...**
- They can't be satisfied simultaneously
  - Trade-off is required (sounds familiar....)
- Common approaches
  - Pick a subset to satisfy => leading to different mechanisms (voting rules)
  - More CS-style approach: allow some properties to be “**approximately**” satisfied
  - Crowdsourcing the trade-offs
    - (then often again leading to new trade-offs....)