Moral Preferences

FRANCESCA ROSSI





Decision making

- Based on our preferences over the options
- Social context: aggregation of the individuals' preferences
 - Voting rules: from collection of preference orderings to a single preference ordering (or its top element)
- Preference modelling and reasoning frameworks
 - CP-nets, UCP-nets, TCP-nets, soft constraints, etc.
- Rationality of individual preferences
 - Preference ordering is transitive
- Desired properties of preference aggregation process and result
 - Unanimity, Pareto optimality, monotonicity, participation, fairness, strategy-proofness, nondictatorship, etc.
- No mention of morality or ethics
 - Rationality does not imply morality
- How to embed morality in a decision process, and to generate moral decisions?



Why moral decision making?

- We need to trust AI systems
- They live and work with us in critical environments
 - They will drive our cars, take care of our elderly people and kids, they suggest diagnosis and therapies
 - Besides suggesting things to buy or posts to read
- Nothing morally wrong should be done
- Autonomous AI system should behave ethically
 - Or we won't let them be autonomous
- In human-machine environments, machine members of the team should be ethical
 - Or teamwork would be precluded because of lack of trust



Why ethics in AI?

• Butler robot

• He should prepare dinner, but should not cook the cat if nothing is in the fridge!

• Self-driving cars

- It should bring us home, but should not run over pedestrians to make us get there at the desired time!
- Companion robot for elderly people
 - It should remind to take medicines, but should also do so in a gentle way
- Healthcare decision support systems
 - They should not suggest a therapy only because it is the least expensive



Preferences

- They usually define a partial order over the options
 - Or total order with ties
- Qualitative or quantitative ways to specify preferences
 - I prefer Breakfast at Tiffany's to Terminator
 - 5 stars to Ex Machina and 2 to Her
- Unacceptable options are ruled out
 - Constraints
- Compact ways to model the preference ordering
 - When options have a combinatorial structure
 - × Combination of features
- Efficient ways to find the most preferred option and to check if an option dominates another one





Preference aggregation

- From the individuals' preferences to a collective decision
- Voting rules
 - Acting over full decisions or features of them
 - Borda, plurality, Copeland, cup rule, approval, k-approval, Kemeny, Single Transferrable Vote, Veto, Minimax, Range, Schulze, Banks, Slater, Bucklin, Dogson, ...
 - Fair, unanimous, monotonic, Condorcet-consistent, neutral, anonimous, ...





Morality and ethics

• Priority over actions

• Based on what is morally right or wrong

• Several ethical theories for humans

- Consequentialism: actions consequences are evaluated in terms of good and bad, and agent should minimize bad and maximizes good
- Deontologism: Actions are predefined as good or bad, agent should choose the best action
- Notion of right and wrong depends on context
 - Ethical theory: function from a context to a partial order over actions
 - Some actions can be incomparable
- Not that different from what preferences define!



An Ethical Spectrum

Research question 1: ethics modelling and reasoning framework

- Are existing preference modeling and reasoning frameworks ready to be used to model and reason with ethics theories?
- Do they need to be adapted?
- Do we need new ones?
- Can we just merge moral and preference orders to generate moral preferences?

Research question 2: moral preferences

- How to combine ethics and preference orderings?
- What properties do we want to assure for the combination?
- Example:
 - two CP-nets (one of the moral order and another one for the preferences)
 - Syntactically and semantically merged
 - Priority to moral order
 - Preferences to dictate only when consistent with ethics theory







Research question 3: Preference/ethics modelling

- Preference elicitation already a very difficult task
- Elicitating the moral ordering seems even more elusive task
- In a social context, people, change their moral attitude over time because of social interaction
- Various approaches to define ethical principles
- Top-down: set of rules to code all possible situations and solutions to ethical dilemmas
 - Works in very narrow domains only
- Bottom-up: learn by observing human behavior
 Could miss basis athies principles
 - Could miss basic ethics principles
- How to combine top-down with bottom-up approaches?
- Do we need more complex approaches?

Research question 4: explanation and correctness

- Machine learning approaches are opaque
- Do not assure correctness or optimality
- How to provide explanation capabilities in ML based systems?
- How to prove that nothing wrong will ever happen?
- Are existing software verification techniques enough?
- Can we generate decision trees that are faithful to the ML system behavior?

Research question 5: Meta-preferences and moral deviation

- Preferences change over time
 - From societal interaction
- Reconciliation of individual preferences with social reason
- Improvement steps: from one preference ordering to a "better" one
 - Need to be able to judge preference orderings
 - o "Morality requires judgment over preferences", Sen 1974
- Metarankins (or metapreferences) to formalize preference modifications
- Moral code: ranking over preference orderings
 Notion of distance to measure the deviation of any action from the moral code
- How to measure the deviation of a collective or individual choice from a moral code?
- Monotonicity of moral preference aggregation
 - If an individual moves to a more moral preference order, the collective choice should be more moral

- Neuroscientists have shown that human moral judgment does not come from a dedicated moral system
- Product of interaction of many brain networks, each working in narrow context
- Is this true also for AI systems?
- Can narrow AI systems be moral?
- Or do we need to build AGI before we can have morality at all?

Summary

• Trusting AI

- Autonomous systems
- Human-machine environments
- Need to make sure they behave morally
- Moral codes and preferences both define priorities over actions
- Need for both preferences and morality in decision making
 - Individual and group decision making