SAFETY CONSTRAINTS AND ETHICAL PRINCIPLES IN COLLECTIVE DECISION MAKING SYSTEMS

Francesca Rossi

UNIVERSITÀ DEGLI STUDI DI PADOVA

Research partners

- Francesca Rossi (University of Padova, Italy)
  - Constraint solving, preference reasoning, computational social choice
- Joshua Greene (Harvard University, USA)
  - Psychology, neuroscience, moral judgement, decision making
- John Tasioulas (King's College London, UK)
  - Philosophy, ethics, morality, human rights
- K. Brent Venable (Tulane University and IHMC, USA)
  - Preference reasoning and aggregation, constraint solving
- Brian Williams (MIT, USA)
  - Model-based programming, cooperative and space robotics, scheduling under uncertainty

Project's goals

- Intelligent agents everywhere, interacting and working with humans
  - Driving, assistive technology, healthcare, financial domain, personal assistants, etc.
- Need for collective decision making
- Also need for trusting intelligent agents in their autonomous decisions (or suggestions for decision)
- Aim:
  - Embedding of safety constraints, moral values, ethical principles, in agents and hybrid agents/human decision making

How we plan to achieve them

- Adapting current (logic-based) modelling and reasoning frameworks
  - Soft constraints, CP-nets, constraint-based scheduling under uncertainty
- Modelling ethical principles
  - Constraints to specify the basic ethical laws, plus prioritized context-dependent constraints over possible actions
  - Conflict resolution engine
- Replacing preference aggregation with constraint/value/ethics/ preference fusion
  - Agents' preferences should be consistent with the systems' safety constraints, the agents' moral values, and the ethical principles of both individual agents and the collective decision making system
- Learning ethical principles
- Predicting possible ethical violation

Appointments

- Napoleon Xanthoulis
  - PhD student in law, King's College London
- Andrea Loreggia
  - Post-doc, University of Padova
- Alex Gain and Kyle Bogosian
  - Undegraduate students, Tulane University
- Zac Bathen
  - Undergraduate student, Tulane University
- Kyle Dillon
  - PhD student psych, Tulane University

Publications

- Embedding Ethical Principles in Collective Decision Support Systems
  - J. Greene, F. Rossi, J. Tasioulas, K. Venable, B. Williams. Proc. AAAI 2016.
  - AAAI Blue Sky Award.
- Our driverless dilemma
  - J. Greene. *Science*, *352*(6293), 1514-1515
  - Cited in New York Times, the Washington Post, the Los Angeles Times, and the Daily Mail.
- Moral preferences
  - F. Rossi, IJCAI 2016 workshop on preferences
- Ethical Preference-Based Decision Support Systems
  - F. Rossi, CONCUR 2016 (paper for invited talk)

Courses and talks

- Courses:
  - Ethics for Artificial Intelligence
    - B. Venable. Tulane University, 2016
  - Evolving Morality: From Primordial Soup to Superintelligent Machines
    - J. Greene, Harvard University, 2016
- Invited talks:
  - Safety constraints and ethical principles in collective decision making systems, F. Rossi, KI 2015
  - Moral Preferences, F. Rossi, ACS 2016
  - Ethical Preference-Based Decision Support Systems, F. Rossi, CONCUR 2016
  - Moral Preferences, F. Rossi, MIRI CSRBAI 2016
  - AI ethics, F. Rossi, TEDx Lake Como 2016
- Other:
  - Critical review on ethical principles in driverless cars
    - J. Tasioulas, 2016
  - Repository of papers on ethics in AI
    - B. Venable, 2016

Ongoing work and future directions

- Identified scenarios
  - self driving cars
  - crowd sourced navigation
- Principles for both scenarios
  - Close to implementation level, rather than philosophical
  - To be adopted by manufacturers
- Mimic human moral judgment
  - Optimization (utilitarianism), hard constraints (deontology), pattern matching (virtue ethics)
  - Compact preference modeling
- Centralized system that control many AVs
  - Waze but with control rather than advise
  - Safety concerns for the entire system
  - Fairness
  - Risk-aware traffic network
  - Game theory rather than voting theory
- Morality as a distance between preference orderings

JAIR track on AI and society

- To understand/educate about the impact of AI on society
  - Both short and long term
- Not just scientific papers
  - Also viewpoints (2000 words), point/counterpoints papers, and multi-position papers
- Subtracks
  - AI and economics
  - AI and law
  - Long term societal impact of AI
  - AI and autonomy
  - AI and philosophy
  - AI and law

- Submit papers!
- And let me know if you need more info

Partnership on AI
www.partnershiponai.org

- Founding members: IBM, Microsoft, Amazon, Facebook, Google/DeepMind
- Board with as many non-corporate members (to be announced very soon)
- Goal: discuss/study/advice/educate/share on best practices for ethical development + ethical behavior of AI systems
- Topics:
  - Trust, privacy, interpretability, explanations, …
  - Ethics: moral values, professional codes, social norms, codes of conduct, …
- Engagement with academia, scientific associations, non-profit organizations, industrial sectors, policy makers, …
- Deliverables: Events, white papers, research efforts, …
- Open licence for all activities
- It is not: regulatory, lobbying, governing body

- If you are interested in joining forces, talk to me (or Yann, or Mustafa, or Demis)

amazon.com DeepMind Google

facebook IBM Microsoft