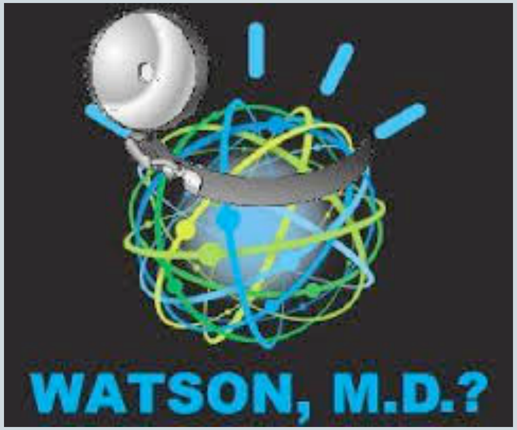
Ethical Intelligent Agents

FRANCESCA ROSSI
UNIVERSITY OF PADOVA

AI is and will be beneficial

- Improving medicine, food production, finance, routine jobs, etc.
- Decreasing cars' accidents, developing assistive technology, solving environmental problems, etc.
- Enabling technology for many kind of workers
  - Doctors, scientists, etc.
  - Too much data for a human to read/assimilate/digest/link



However

- Common sense, unstated assumptions need to be spelled out when stating the goal of an AI
  - Otherwise the wrong goal may be achieved
- AI cannot start thinking on its own, as some fear
  - But it can maximize a "wrong" objective function, if we are not careful
- We need autonomy (although it does not coincide with intelligence)
  - If we want to fully exploit the AI's capabilities
  - But autonomous agents need to be trusted
    - Explanation capabilities
- Narrow domains for many current AIs
  - But still can have huge impact
- Full AI very improbable to be achieved anytime soon
  - The usual 20 years prediction is just a number
  - Common sense reasoning needs to be understood and coded
  - Deep learning does not solve everything

Ethical group decision making systems

- Autonomous agents everywhere, interacting and working with humans
  - Driving, assistive technology, healthcare, etc.
- Need for collective decision making
- Need to trust intelligent agents in their autonomous decisions
- Embedding 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

Research group: CS/AI/philosohy/psychology (FLI funding)

Open letters

- Making AI beneficial (January 2015)
  - Constructive approach, to avoid extreme positions in the debate
  - 37 new research projects, based on solid scientific grounds
- Autonomous weapons (July 2015)
  - Is a ban appropriate for AI research directly intended for such a purpose?
  - Many examples in other research communities
    - Medicine, cryptology, physics, chemistry, molecular biology, psychology
- Role of AI associations (AAAI, IJCAI, …)
  - Inform members and

  work with agencies/governments/international bodies

future of life INSTITUTE

IJCAI International Joint Conferences on Artificial Intelligence