

Approaches and Architectures:

Bottom-Up vs. *Top-Down*

Christopher Mühl

Seminar: Normative Reasoning and Machine
Ethics (Prof. Benz Müller, FU Berlin WiSe 19/20)

Structure

1. What is Machine Ethics?
2. Categories of Approaches
3. Comparison: Advantages and Disadvantages
4. Examples:
 - Bottom-Up: CIRL
 - Top-Down: LOGIKEY
5. Discussion

What is Machine Ethics?

1. What is Machine Ethics?

Landscape of research field

- machine ethics is branch of AI ethics
- AI ethics: minimize ethical harms by AI
- machine ethics: create ethical machines

Ethical agencies

1. ethical impact agents
2. implicit ethical agents
3. explicit ethical agents
4. full ethical agents

Categories of Approaches

2. Categories of Approaches

Bottom-Up Approach

- engineering sense:
 - use performance measure
- ethical sense:
 - treat normative values as being implicit in activity of agents
 - not explicitly articulated (or even articulable)
- merged sense:
 - create agents with accurate understanding of own and other's morality

2. Categories of Approaches

Top-Down Approach

- engineering sense:
 - decompose task into subtasks
- ethical sense:
 - take general ethical theory
 - derive consequences for particular cases
- merged sense:
 - take ethical theory
 - analyze requirements
 - design subsystems to implement that theory

2. Categories of Approaches

Hybrid Approach

- merge bottom-up and top-down
 - bottom-up for socialization
 - top-down for basic norms

Different Approaches for specific skills

- supervised learning: deontological rules
- reinforcement learning: socialization
- applied game theory: multiple-party interests
- probabilistic programming: uncertainty management
- scenario rendering: modelling interactions
- inverse reinforcement learning: modelling intent

Comparison:

Advantages and Disadvantages

3. Comparison:

Advantages and Disadvantages

Bottom-Up

- advantages:
 - assembling components to achieve goal
 - no explicit rules are needed
 - step-by-step learning
- disadvantages:
 - not transparent
 - bias-prone
 - insufficient artificial environments for training
- problems:
 - does morality emerge from assembly?
 - how to verify morality?

3. Comparison:

Advantages and Disadvantages

Top-Down

- advantages:
 - transparent
 - easy fitting to new contexts by adding rules
- disadvantages:
 - how to formulate adequate universal rules that include unknown scenarios?
 - trade-off between vagueness and inflexibility
- problems:
 - which rules to choose?
 - consequentialist
 - deontology
 - virtue ethics
 - what to decide when several rules conflict?
 - when to terminate calculation of best action?

Bottom-Up Example:

CIRL

4.1 Bottom-Up Example:

CIRL

Cooperative Inverse Reinforcement Learning

- *implicit ethical agent*
- robot observes human
- estimates human's intent to act accordingly

Problems

- robot should not copy all behaviors
 - constraint reward function to optimize reward for human
- observing actions is inefficient
 - actively teach and ask instead

4.1 Bottom-Up Example: CIRL

Definition (simplified)

- two-player Markov game (human H , robot R)
- each timestep t : H and R observe state s_t and select action a_t^H, a_t^R
- both receive reward r_t and update behavior π^H, π^R
- behavior is function of observation history that determines action selection

Top-down Example:

LOGIKEY

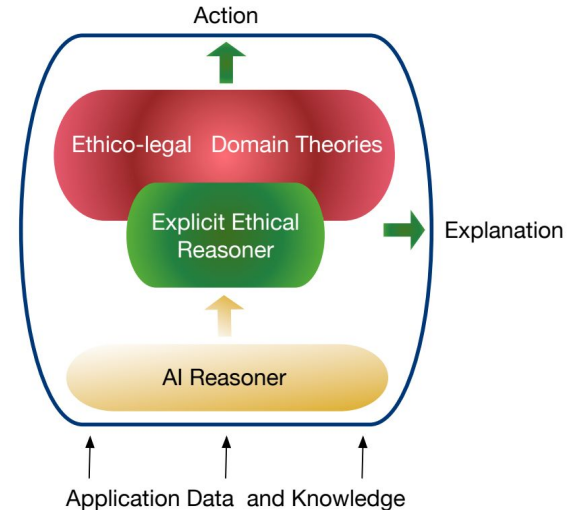
4.2 Top-Down Example:

LOGIKEY

Logic and Knowledge Engineering Framework and Methodology

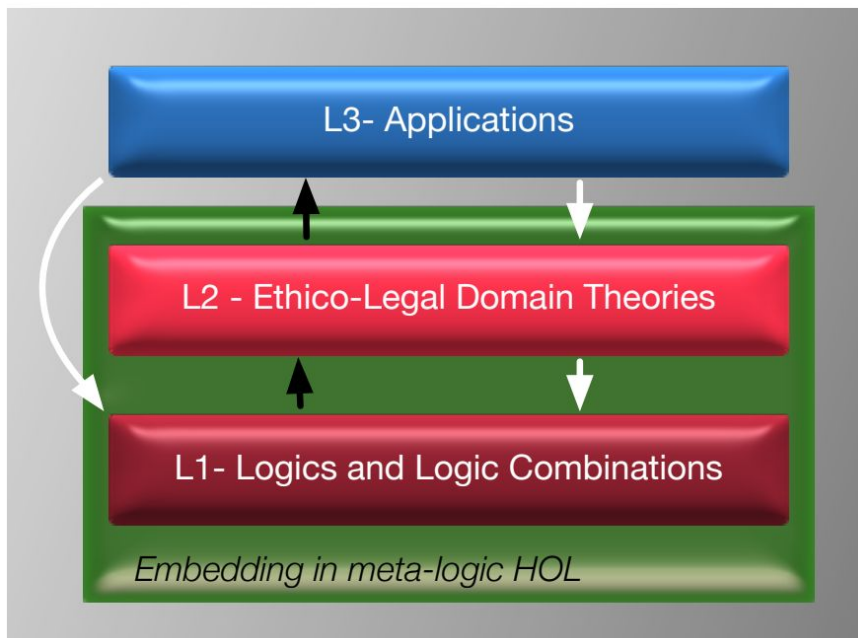
- (part of) *explicit ethical agent*
- framework for normative reasoning tools
- uses higher order logic (HOL)

Scheme



4.2 Top-Down Example: LOGIKEY

- LOGIKEY comprises 3 layers
- in each layer the concrete content is interchangeable
- L1: Logics
 - modal logic
 - norm-based
- L2: Ethical Theories
 - standard deontic logic (SDL)
 - dyadic deontic logic (DDL)
 - input/output (I/O) logic
- L3: Applications
 - General Data Protection Regulation (GDPR)
 - Gewirth's Principle of Generic Consistency



Discussion

5. Discussion

1. What skills are needed for *full ethical agents*?
2. How to approach them?

References

1. W. Wallach, C. Allen, I. Smit. (2008). Machine Morality: Bottom-up and Top-down Approaches for Modeling Human Moral Faculties.
2. C. Benz Müller, X. Parent, L. van der Torre. (2019). Designing Normative Theories of Ethical Reasoning: Formal Framework, Methodology, and Tool Support.
3. D. Hadfield-Menell, A. Dragan, P. Abbeel, S. Russell. (2016). Cooperative Inverse Reinforcement Learning.
4. A. F. Winfield, K. Michael, J. Pitt and V. Evers. (2019). Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems [Scanning the Issue].
5. <https://www.ethicsnet.org/blog/Blog/approaches-to-ai-values>

Approaches and Architectures:

Bottom-Up vs. *Top-Down*

Christopher Mühl

Seminar: Normative Reasoning and Machine
Ethics (Prof. Benz Müller, FU Berlin WiSe 19/20)