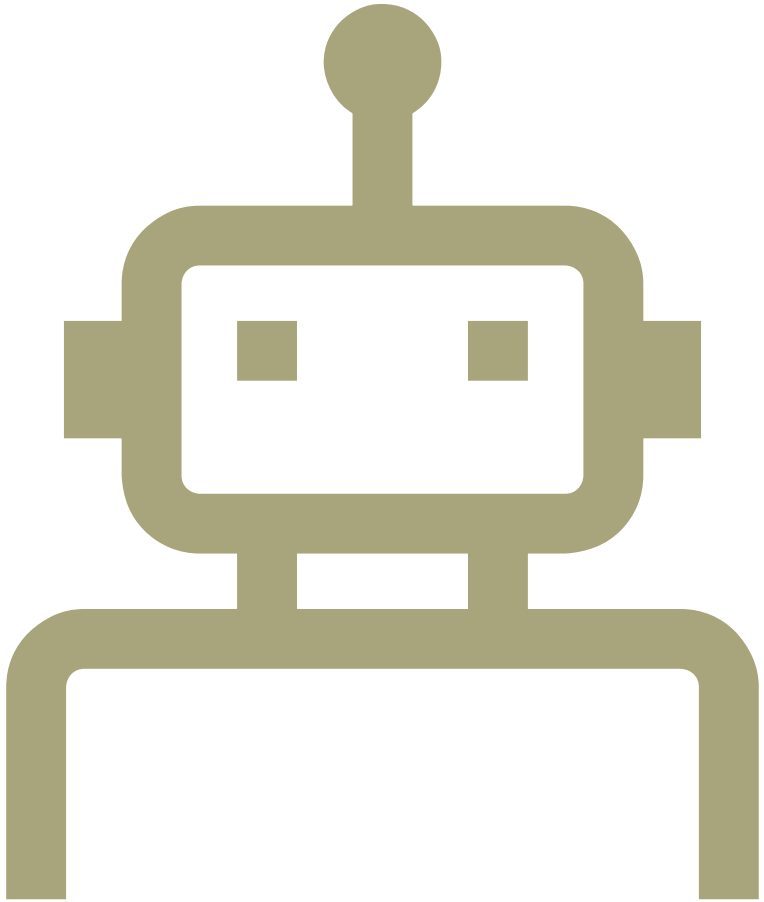
Machine Ethics

Why is it challenging?

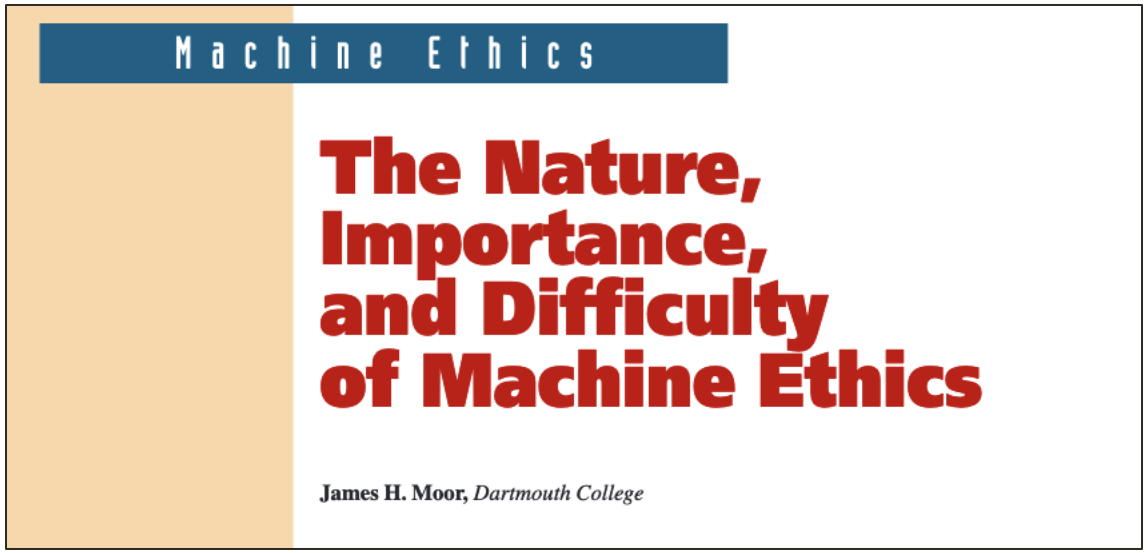


What is AI? (an agent-based definition)

- As per Poole and Mackworth (2017) "Artificial Intelligence is the field that studies the synthesis and analysis of computational agents that act intelligently".
- Agent = an entity that acts in an environment
- Computational agent = an agent whose decisions about its actions can be explained in terms of computation
- We will look at how computational agents could make ethical decisions.

Machine Ethics

- How to automate moral reasoning for computational agents?



* Moor, James H.: The Nature, Importance, and Difficulty of Machine Ethics. In: IEEE Intelligent Systems (2006), Juli, Nr. 4, S. 18–21.

Humans are machines and humans have ethics.

Machine ethics does not exist because ethics is simply emotional

Could a computer operate ethically because it is internally ethical in some way?

* *Moor, James H.: The Nature, Importance, and Difficulty of Machine Ethics. In: IEEE Intelligent Systems (2006), Juli, Nr. 4, S. 18–21.*

Machine Ethics -- Ethical Agents

Ethical-impact agents: Designing a machine solution for a specific task, which impacts ethical issues. (ex: loan system)

Implicit ethical agents: Constraining the machine's actions to avoid unethical outcomes. (ex: banking agents)

Explicit ethical agents: Representing ethics explicitly. (ex: modeling privacy preferences as logic-based rules)

Full ethical agents: Making judgments with justifications while having features such as consciousness, intentionality and free will.

Developing Explicit Ethical Agents

- They fall short of being full ethical agents, BUT they could prevent help prevent unethical outcomes.
- Why is Machine Ethics important?
  - We want machines to treat us well!
  - Future machines will likely have increased control and autonomy. They will need more powerful machine ethics.
  - We should also understand ethics. Programming or teaching a machine to make ethical decisions is also good for us!

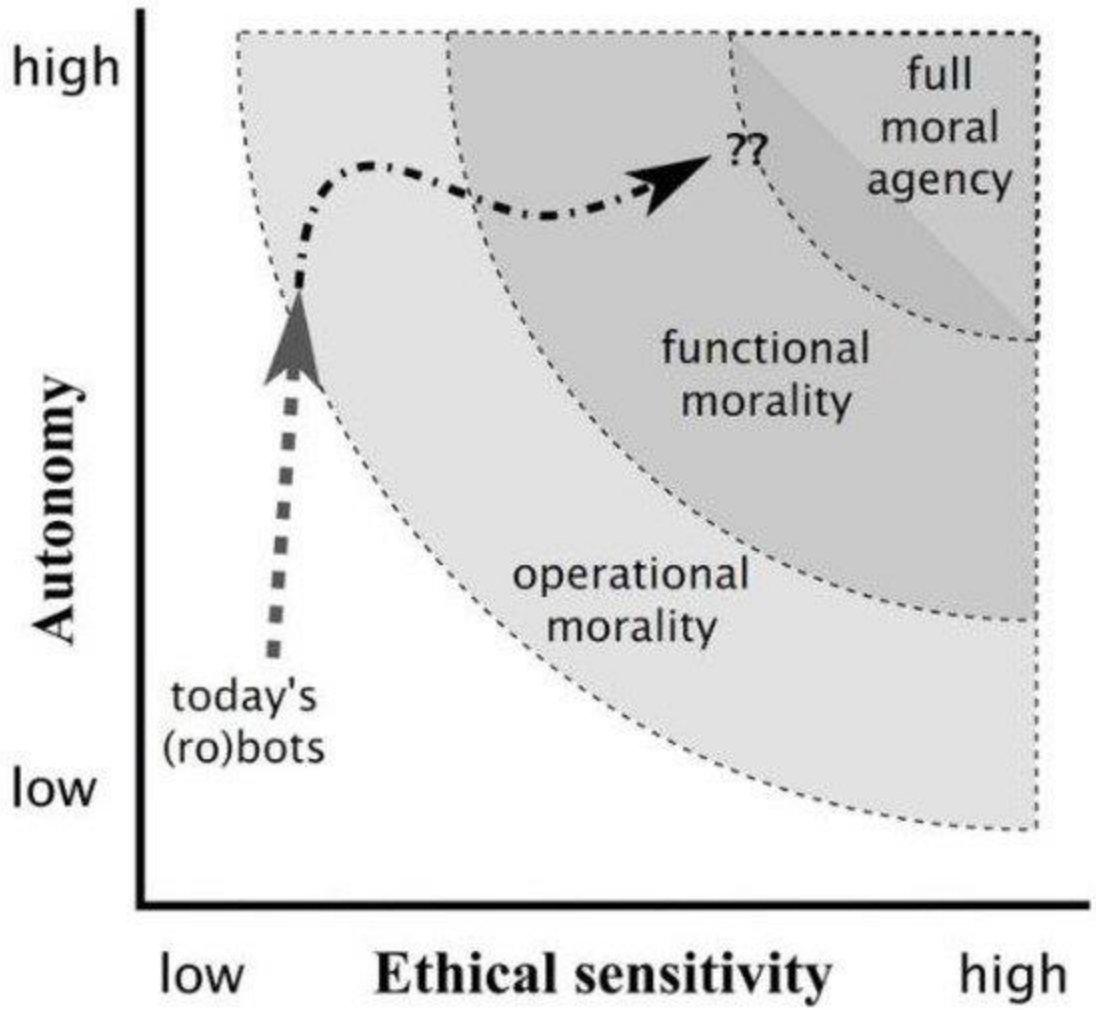
** Moor, James H.: The Nature, Importance, and Difficulty of Machine Ethics. In: IEEE Intelligent Systems (2006), Juli, Nr. 4, S. 18–21.*

Why is Machine Ethics a "myth"?

- We have a **limited understanding** of ethical theories.
  - Disagreement on the subject
  - Conflicting ethical intuitions and beliefs
  - Different than programming an agent to do some complex task where moves are well defined (e.g., chess)
- We need to **understand learning** better (e.g., machine learning etc.)
- Computers have **limited commonsense knowledge**.

** Moor, James H.: The Nature, Importance, and Difficulty of Machine Ethics. In: IEEE Intelligent Systems (2006), Juli, Nr. 4, S. 18–21.*

Wallach and Allen Approach for Categorization of Machine Ethics



Wallach and Allen. 2008. Moral machines: Teaching robots right from wrong. Oxford University Press, Oxford, UK.

Wallach and Allen Approach for Categorization of Machine Ethics

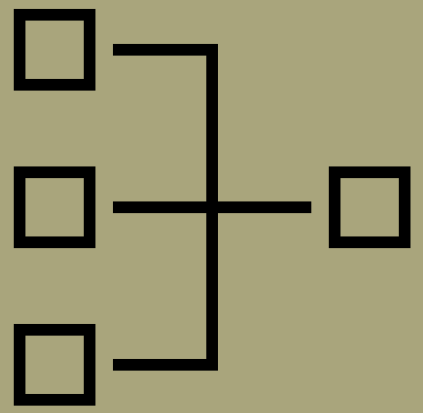
- Top-Down
  - Start with an ethical theory, identify smaller problems and solve them.
  - Pros: no need to identify additional problems
  - Cons: Not clear from the beginning if subproblems are solvable
- Bottom-Up
  - Start with data, and learn ethical behavior from data.
  - Pros: Subproblems are solvable
  - Cons: Non-necessary subproblems may be dealt with.

Wallach and Allen. 2008. Moral machines: Teaching robots right from wrong. Oxford University Press, Oxford, UK.

Louise Dennis Approach for Categorization of Ethical Systems

- **Constraint-Based** Ethical Systems
  - Ethics is placed on some sub-system that guides/constrains the actions of other parts.
  - Other parts of the system can guide the decision-making process of the agents.

- **Global** Ethical Systems
  - All decisions are ethical.



How to build **ethical** systems?

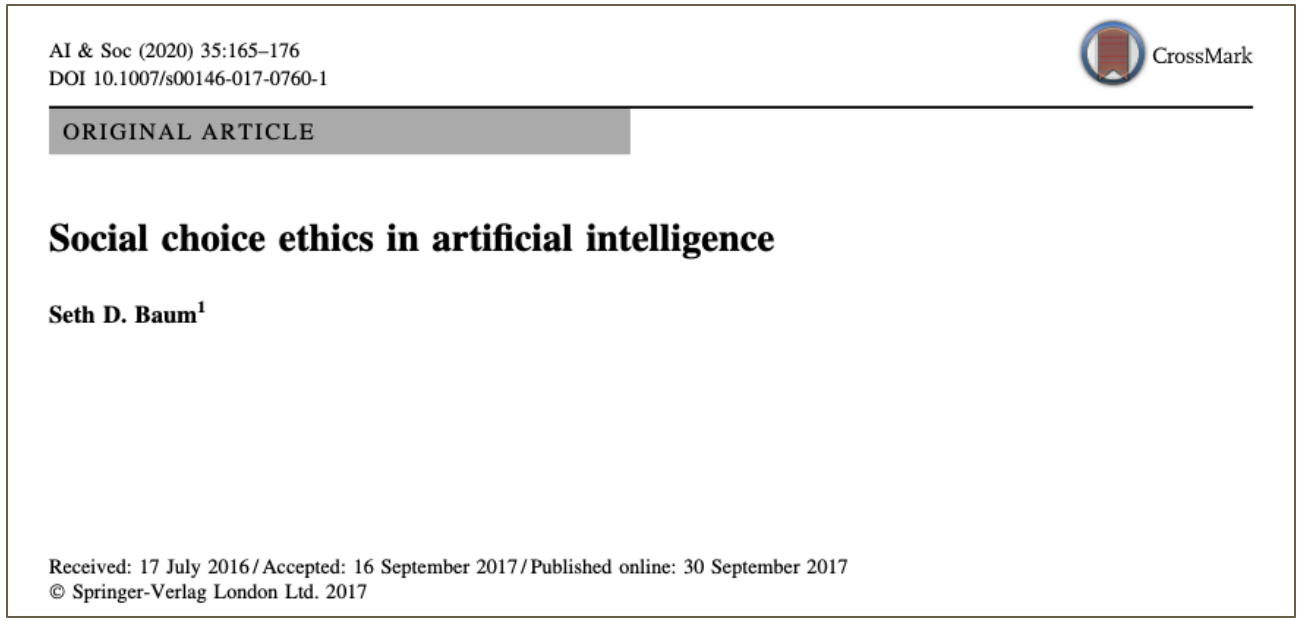
Social Choice and Machine Ethics

- We often talk about implementing values or obligations.
- We are now interested in the question of whose values/obligations a machine should implement.
- Once we know what we want to implement, we can develop algorithms to verify machine ethics systems (e.g., Isabelle).

Consequentialist Theories (revisited)

- Ethical Egoism
  - Focuses on own best interests
- Utilitarianism
  - Focuses on everyone
  - Act-utilitarianism:
    - from individual to society
  - Rule-utilitarianism:
    - A rule to follow to achieve overall good

Social Choice Ethics in AI



AI & Soc (2020) 35:165–176
DOI 10.1007/s00146-017-0760-1

CrossMark

ORIGINAL ARTICLE

Social choice ethics in artificial intelligence

Seth D. Baum[1]

Received: 17 July 2016 / Accepted: 16 September 2017 / Published online: 30 September 2017

Seth D. Baum. 2020. Social choice ethics in artificial intelligence. AI & Society 35, 1 (2020), 165–176.

Social Choice Ethics in AI

- Goal:
  Designing AI to act according to the aggregate views of society (i.e., bottom-up).
- Value-based decision making:
  - Standing (whose ethics views)
  - Measurement (identifying views)
  - Aggregation (combining to a single view)
- Non-social ethics could be even more challenging
  - Considering future generations, or the AI itself

Seth D. Baum. 2020. Social choice ethics in artificial intelligence. AI & Society 35, 1 (2020), 165–176.

Summary

- Machine Ethics is:
  - a way to realize Normative Ethics and Applied Ethics together.
- Many categorization systems exist:
  - Moore's Ethical Agents, Wallach and Allen, Louise Dennis …
- Social Choice theory is:
  - looking at the problem of understanding values/obligations of a society