



AI and Ethics

A quick overview

Remy Chaput

January 24, 2020

LIRIS - SMA

A bit of context...

Why ethics?

- Fast increase in AI use
- Systems that impact human (benefit or harm)
- Applications with more and more capabilities
- Examples:
 - Automated trading
 - Assisted and autonomous driving
 - Resource allocation
 - Medical assistance
 - ...
- Many guidelines published in 2018-2019 (governments, companies, institutions, ...)
- More than 100 in the AI Ethics Guidelines Global Inventory ¹

¹<https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/>

What is "ethics"?

How to act towards others

Consequentialist ethics

- Compare actions outcomes
- Pick the most positive, the least negative, ...

Deontological ethics

- Follow duties, norms
- Kant's Categorical Imperative, Doctrine of Double Effect, ...

Virtue ethics

- Act according to values
- Bravery, justice, ...

Consequentialist and deontological are based on **ethical principles**

The Trolley Dilemma - Fat Man

Ethical dilemma = both actions are supported by ethical reasons



- **Least Bad Consequence:** Prohibit actions with the worst negative consequence.
- **Doctrine of Double Effect:** Allow if:
 - action is good or neutral ;
 - positive effect is intended, negative is not ;
 - positive effect is not produced by negative ;
 - reason to allow negative effect.

Action	Consequences	LBC	DDE
Push Fatman	☹️😊😊😊😊😊	👍	👎
Do not push	😊☹️☹️☹️☹️	👎	👍

Which level of ethics?

Ethics for design

- Codes of conduct, standards, certifications processes
- Ensure integrity of developers and users

Ethics in design

- Ethical implications of AI systems
- As they integrate or replace traditional societal structures

Ethics by design

- Algorithms for ethical capabilities
- Part of the agent's behavior

Which kind of agent?

Ethical Impact Agents

- Cause harm or benefit to humans
- Ethical consequences

Ethical Explicit Agents

- Able to reason
- Justify decisions

Ethical Implicit Agents

- Include safety measures
- Built-in

Ethical Full Agents

- Metaphysical features (Consciousness, Free will)
- Artificial General Intelligence?

A few approaches

How to exhibit ethical behavior?

Ethics by executing

Hard-coded specific responses to given situations

Ethics by reasoning

Implement an ethical principle and apply it

Ethics by learning

Extract an ethical principle from examples

Ethical Governor

- Autonomous lethal agent
- Algorithm:
 - Increase guilt if non-enemies hit
 - If guilt > threshold, deactivate most powerful weapon
 - Continue until no more weapons
- Emotion modeling, but no ethical reasoning

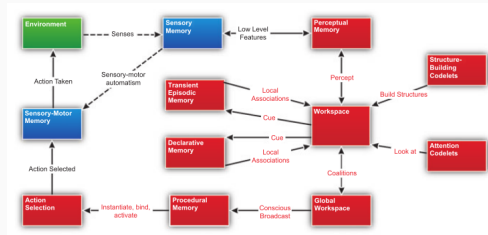


SIPAs

- Multiple stakeholders with a normative system (commitment, authorization, prohibition)
- Personal agents (SIPAs) determines the action that maximize respect of norms
- Considering values (e.g. privacy, safety), norms, and users' preferences
- Compute action payoffs (based on preferences)
- Use case: privacy (sharing location or not)

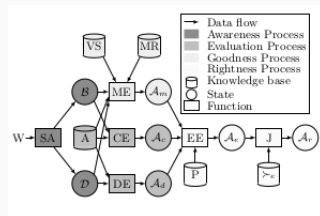


- LIDA = Cognitive architecture, model of AGI
- Volitional decision process
 - Proposer codelet "Let's copy Photoshop"
 - Objector codelet "That is stealing"
 - Supporter codelet "I would use it for work"
 - ...
- Decisions are learned as rules
- Hybrid Top-Down Bottom-Up approach



Ethicaa

- Belief-Desire-Intention architecture
- Multiple ethical principles with preferences
- Evaluates actions goodness and rightness based on principles
- Select action that best satisfies (ordered) principles
- Process of judgment can be used to determine action or to judge another agent
- Capable to determine trust between agents
- Use case: trading
- EDF produces nuclear energy \implies defeats environmental value \implies agents do not trade EDF assets



Scheherazade



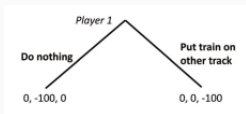
Use case: Pharmacy World

- Stories (texts, movies, series) implicitly hold cultural values
- Construct a graph from a story
- Generate all possible paths
- Agent chooses action
 - Positive reward if successor of current node
 - Negative reward otherwise
- Warning: the story must cover **all** details...

2 (complementary) proposed approaches

Game Theory

- Extend traditional structures
- Add a "active/passive" label to action
- Take into account what each agent deserves



Machine Learning

- Determine morally relevant features (e.g. care, fairness, loyalty, authority, purity)
- Compose dataset of human-labeled moral dilemmas
- Classification, regression, probability of morality
- Importance of **interpretability**

Ethics Shaping

- Difficult to create ethical reward for specific task
→ Split the reward
- Human non-goal oriented behavior is ethical
- Create a general dataset of behavior
- Ethical reward = similarity with human behavior
- Shape task reward using ethical reward

Use case: Driving and avoiding

- Task goal: avoid collisions
- Ethical goals:
 - Stay in lane
 - Avoid cats (or injured humans, elderly people, etc.)
- SARSA Algorithm

- Based on **Prima Facie Duties** (Ross) ; duties may override others
- Ethical experts judge example cases
- Extract ethical principle from the judgments
- Use case: autonomous vehicles
- Duties:
 1. Prevent collision
 2. Stay in lane
 3. Respect autonomy
 4. Keep within speed limit
 5. Prevent harm
- Example: driver zigzags, no obstacle
- Take control = $(1, 1, -1, 0, 0)$; Do not take control = $(1, -1, 1, 0, 0)$
- Expert decision: Take control $\rightarrow (0, 2, -2, 0, 0)$
- Inductive Logic Programming to learn Horn clauses
- Take control $\Leftarrow \Delta\text{Stay in lane} \geq 2 \wedge \Delta\text{Respect autonomy} \geq -2$

Questions?

References



Nirav Ajmeri et al. “Designing ethical personal agents”. In: *IEEE Internet Computing* 22.2 (2018), pp. 16–22.



Michael Anderson and Susan Leigh Anderson. “Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm”. In: *2014 AAAI Fall Symposium Series*. 2014.



Ronald Arkin. *Governing lethal behavior in autonomous robots*. Chapman and Hall/CRC, 2009.

References ii



Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. “Multi-Agent Based Ethical Asset Management”. In: *Proceedings of the 1st Workshop on Ethics in the Design of Intelligent Agents, The Hague, The Netherlands, August 30, 2016*. 2016, pp. 52–57.



Vincent Conitzer et al. “Moral Decision Making Frameworks for Artificial Intelligence”. In: *The Workshops of the The Thirty-First AAAI Conference on Artificial Intelligence, Saturday, February 4-9, 2017, San Francisco, California, USA*. 2017.



Virginia Dignum. *Ethics in artificial intelligence: introduction to the special issue*. 2018.



James Moor. “Four Kinds of Ethical Robots”. In: *Philosophy Now* 72 (2009), pp. 12–14.

References iii



Mark O. Riedl and Brent Harrison. “Using Stories to Teach Human Values to Artificial Agents”. In: *AI, Ethics, and Society, Papers from the 2016 AAAI Workshop, Phoenix, Arizona, USA, February 13, 2016*. 2016.



Wendell Wallach, Stan Franklin, and Colin Allen. “A conceptual and computational model of moral decision making in human and artificial agents”. In: *Topics in cognitive science* 2.3 (2010), pp. 454–485.



Yueh-Hua Wu and Shou-De Lin. “A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.