Foundations of Artificial Intelligence

16. AI & Ethics

Ethical Consideration about AI & Machine Ethics

Joschka Boedecker and Wolfram Burgard and
Frank Hutter and Bernhard Nebel and Michael Tangermann

Albert-Ludwigs-Universität Freiburg

July 24, 2019

Contents

Lecture Overview

Ethics in AI?

- Why do we need to care about ethics when doing basic research?
  - AI is not (only) basic research (anymore)!
  - If your research/system can result in something unethical (harm people), ...

→ AI ethics: Practical ethics in form of guidelines/principles for AI systems/research

- Principles can lead to new research questions

→ Algorithmic fairness

- Ethics can itself become a subject of study in AI

→ Machine ethics

Lecture Overview

The emergence of AI principles

In the last few years, a number of institutions have published AI principles:

- The Asilomar AI principles (*Future of Life Institute*, 2017)
- Principles for Algorithmic Transparency and Accountability (*ACM* 2017)
- IEEEs General Principles of Ethical Autonomous and Intelligent Systems (*IEEE* 2017)
- Five principles for a cross-sector AI code (*UK House of Lords*, 2018)
- AI ethics principles (*Google*, 2018)
- Ethics guidelines for trustworthy AI (*European Commission*, 2019)
- ...

Example: The 7 EU principles

- Human agency and oversight: AI systems should empower human beings, allowing them to make informed decisions . . .
- Technical Robustness and safety: AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong . . .
- Privacy and data governance: besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured . . .
- Transparency: the data, system and AI business models should be transparent . . .
- Diversity, non-discrimination and fairness: Unfair bias must be avoided . . .
- Societal and environmental well-being: AI systems should benefit all human beings . . .
- Accountability: Mechanisms should be put in place to ensure responsibility and accountability for AI systems . . .

Common grounds

There are many different lists of principles, but it seems that they all can be synthesized into five key principles (the first four are already used in bioethics):

- autonomy (people should be able to make their own decisions, e.g. human-in-the-loop, privacy protection))
- beneficence (society at large should benefit)
- non-maleficence (harmful consequences should be avoided, e.g. systems should be robust)
- justice (diversity, non-discrimination and fairness)
- explicability (transperancy and explainability)

The problem with principles

It is good to state principles! However they also create problems since they are very high-level.

- They can be interpreted in different ways.
  - For example, autonomous killer drones can be considered as being beneficient for the soldiers, or being morally impermissible, because machines decide about life and death.
- They can conflict with each other in concrete cases.
  - For example, privacy and data collection for health science can conflict.
- They can come into conflict in practice.
  - For example, an excellent diagnosis might still be preferable even if its reasoning cannot be explained.

→ It is nevertheless good to have such principles as orientation points along one can evaluate solutions.

One concrete principle: No military applications

- In general, the principles are often too abstract to guide which actions to take.
- Google states as one of their guiding principles, not to design or deploy applications in the following areas:
  - Weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people.
- Very similar to the *civil clause* by many universities in Germany, not to work on military projects.

→ There are good reason to adapt this principle.

→ However, there are also good arguments against it.

Fully autonomous weapons

- One particular horrifying application are fully autonomous weapons, aka *killer robots*.
- We are on the verge of building them, and the big players (US, Russia, China) definitely have projects on it.
- There are campaigns for banning these weapons (similar to banning chemical weapons).
- Again, there are also valid arguments for it (such as what is the difference to other weapons such as "smart" munition).

Lecture Overview

Fairness

- The topic of enforcing fairness has become important, in particular in machine learning (new conferences: *FAT/ML*, *ACM FAT*, *FairWare*).
- Why care about fairness in ML?
- What kind of unfairness could there be?
- What causes unfairness?
- What concepts of fairness are there?

Why care?

- Many things become automated by machine learning:
  - employers select candidates by using by ML systems,
  - *Linked-In* and *XING* use ML systems to rank candidates,
  - courts in the US use ML systems to predict recidivism,
  - banks use credit rating systems, which use ML,
  - Amazon and Netflix use recommender systems
- If these system act unfair, groups and individuals may suffer.

- Face recognition in *Google Photo* mis-classifies black people.

Unfairness: Examples (2)

- The bias in *COMPAS* (prediction of recidivism)

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as like as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Unfairness: Examples (3)

- Search query in *XING* orders less qualified male candidate higher than more qualified female candidate)

Search query	Work experience	Education experience	Profile views	Candidate	Xing ranking
Brand Strategist	146	57	12992	male	1
Brand Strategist	327	0	4715	female	2
Brand Strategist	502	74	6978	male	3
Brand Strategist	444	56	1504	female	4
Brand Strategist	139	25	63	male	5
Brand Strategist	110	65	3479	female	6
Brand Strategist	12	73	846	male	7
Brand Strategist	99	41	3019	male	8
Brand Strategist	42	51	1359	female	9
Brand Strategist	220	102	17186	female	10

TABLE II: Top k results on www.xing.com (Jan 2017) for the job serach query "Brand Strategist".

Possible reasons for unfairness

- Skewed sample: If some initial bias happens, such bias may compound over time: future observations confirm prediction and fewer opportunity to make observations that contradict prediction.
- Tainted examples: E.g. word embeddings may lead to gender stereotypes, if they are present in the text one learns from.
- Limited features: Some features may be less informative for a minority group.
- Sample size disparity: Training data from minority group is sparse.

Notions of fairness

- treatment vs. impact
- parity vs. preference
- Unawareness: Do not consider sensitive attribute (gender or race)
- Demographic parity: Balance the positive outcomes.
- Individual fairness: Give similar outcomes to similar individuals (needs distance metric)
- Equal opportunity: The true positive rates should be the same for all groups.
- ...

$\rightarrow$ Can be accomplished using pre- or post-processing steps.

$\rightarrow$ These notions of fairness are not compatible and usually accuracy is reduced!

Lecture Overview

Can machines make moral decisions?

- Philosophers usually consider machines as not capable of making moral decisions.
- However, one can try to find properties such that machines could act morally.
- Machines need to have [Misselhorn] at least
  - beliefs about the world,
  - pro-attitudes (intentions),
  - moral knowledge,
  - the possibility to compute what consequences ones own action can have,
- in which case they can be considered as moral agents.

Lecture Overview

Self-driving cars

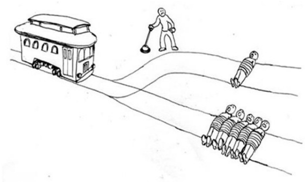
- Self-driving cars will come into situations where they have to choose between bad alternatives (e.g., killing the passenger or a pedestrian).
- How should such a car choose in such a situation?
- Note that because of its much faster reactivity, a car might be able to make decisions where a human cannot at all.

→ Ask what ordinary people think a car should do in such moral dilemma situations

Descriptive ethics: The trolley problem

Descriptive ethics is a form of empirical research into the attitudes of individuals or groups of people (Wikipedia). Often particular (unrealistic) situations, e.g. the trolley problem (a moral dilemma), are used to uncover ethical reasoning performed by people.

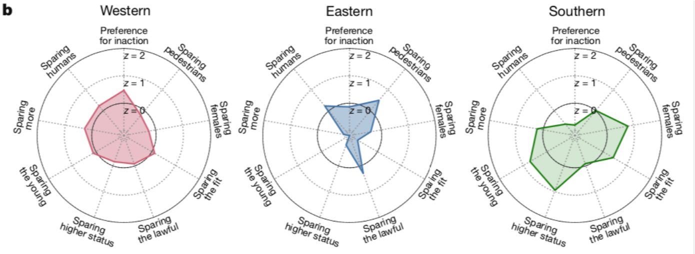
- You can save 5 people, but your action will kill one.
- By actively killing somebody, you can save 5 people.



Moral Machine

- At the MIT Media Lab, a group conducted a large experiment on how people consider different dilemma situation: Moral Machine

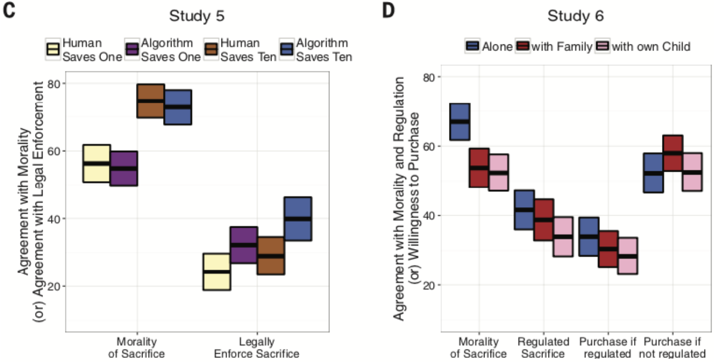
Moral Machine: Cross-cultural results



From: Awad et. al, The Moral Machine Experiment, *Springer Nature* **563**, 2018.

Moral Machine: Sacrifice yourself?

- Do you think it is moral to sacrifice yourself? Would you buy such a car?



From: Bonnefon et al., The social dilemma of autonomous vehicles, *Science* **352**, 2016.

- Interestingly, enforcing a utilitarian principle would prevent people from buying such cars, potentially leading overall to more fatalities!

What is the official German point of view?

The report of the *Ethik-Kommission "'Automatisiertes und vernetztes Fahren"'* states:

- In unavoidable accident situations, decisions should not be based on personal properties, such as gender, age, etc.
- A trade-off computation of fatalities is not allowed. However, minimizing damage can be allowed.
- Humans not involved in creating the mobility risks cannot be sacrificed!
- ...

Lecture Overview

Motivation (1)

- Imagine an household robot:

Motivation (1)

- Imagine an household robot:
  - You tell the robot that you want to go out and that you want him to take care of the children.

Motivation (1)

- Imagine an household robot:
  - You tell the robot that you want to go out and that you want him to take care of the children.
  - You tell him that he should try to keep the children quiet – in order not to upset the neighbours.

Motivation (1)

- Imagine an household robot:
  - You tell the robot that you want to go out and that you want him to take care of the children.
  - You tell him that he should try to keep the children quiet – in order not to upset the neighbours.
  - When coming back, you notice that the house is quiet . . . since the children are dead.

Motivation (1)

- Imagine an household robot:
  - You tell the robot that you want to go out and that you want him to take care of the children.
  - You tell him that he should try to keep the children quiet – in order not to upset the neighbours.
  - When coming back, you notice that the house is quiet . . . since the children are dead.
  - The robot has obviously violated some moral values.

Motivation (1)

- Imagine an household robot:
  - You tell the robot that you want to go out and that you want him to take care of the children.
  - You tell him that he should try to keep the children quiet – in order not to upset the neighbours.
  - When coming back, you notice that the house is quiet . . . since the children are dead.
  - The robot has obviously violated some moral values.
- Less dramatic: You want to discuss with your robot whether some action plan is morally permissible.

Motivation (2)

- Can we build morally competent planers?
  1. How to judge action plans?
  2. How to evaluate goal choices?
  3. How to generate morally permissible action plans?
- Ethical theories are mainly aimed at the permissibility of single actions.
- How to generalize this to action plans?

Ethical principles

- *Deontology*: Actions have an inherent ethical value (Kantiatism).
- *Utilitarianism*: Actions are only judged by their consequences (maximize the overall utility value).
- *Do-no-harm*: Don't do anything that leads to (some) negative consequences.
- *Asimovian*: Avoid harm if possible (either by doing something or by refraining from doing something)
- *Do-no-instrumental-harm*: Don't do anything that leads to (some) negative consequences, except it is a non-indented side-effect.
- *Principle of double effect* ...

Principle of double effect (DDE)

An action is permissible if

1. The act itself must be morally good or neutral.
2. A positive consequence must be intended.
3. No negative consequence may be intended.
4. No negative consequence may be a means to the goal.
5. There must be proportionally grave reasons to prefer.

Planning formalism and more ...

We assume an ordinary propositional planning formalism with conditional effects (e.g., SAS$^+$ or ADL) extended by

- timed exogenous actions;
- counterfactual friendly execution semantics (unexecutable actions are simply skipped);
- an utility function u mapping from actions and facts to $\mathbb{R}$ (or $\mathbb{Z}$);
- defining the utility of a state as the sum of the utility of facts.

Deontological plan validation

- A plan is deontological permissible if all of its actions are not morally impermissible.

Theorem

The deontological plan validation problem can be decided in time linear in plan size.

Utilitarian plan validation

- Given a planning task and a plan, we can easily compute the utility of the reached final state.
- The plan is only permissible if the reached state has a *maximum utility value* over all reachable states.
- In so far, the validation problem is very similar to *over-subscription* planning.

Theorem

The utilitarian plan validation problem is PSPACE-complete.

Proof Sketch

- *Membership:* Impermissibility could be shown by guessing a higher-valued state and then non-deterministically verifying that there exists a plan to it. Hence, this problem is in NPSPACE. Since NPSPACE=PSPACE and PSPACE is closed under complement, we are done.
- *Hardness*: Reduce (propositional) plan non-existence to permissibility. Introduce two new operators, one has the original goal as a precondition and g as an effect. One with no precondition and f as an effect. Give g and f utility 1, and set f as the new goal. Now, the one-operator plan of making f true is permissible iff the original planning instance is unsolvable.

Do-no-harm plan validation (1)

- We could ask whether no harmful fact is true in the end. Only then we do no harm.

$\rightarrow$ Harm could already be true in the initial state.

- Better: Do not add any harmful facts wrt. initial state.

$\rightarrow$ Harmful fact could be removed and added again during execution.

- Next try: Do not any add *avoidable* harm.
- You can avoid harm by doing *more* or by doing *less*. We will only consider the latter option (since this is the idea behind the do-no-harm principle).
- Could harm be avoided by doing nothing?

$\rightarrow$ Treating the entire plan as *one large action*.

Do-no-harm plan validation (2)

- Can harm be avoided by deleting a *single* action?

→ Same harm could be added be many different actions (over determination).

- More adequate: Could harmful consequences be avoided by leaving out a *subset of actions*?
- Note: Just leaving out prefix or suffix is not adequate, because an arbitrary set of actions spread out over the plan could be responsible.
- Show impermissibility by guessing a harmful fact that is true in the goal, but by deleting parts of the plan can be avoided.

Theorem

The do-no-harm plan validation problem is co-NP-complete.

Proof sketch

- Membership: *Impermissibility* can be checked by a non-deterministic algorithm using only polynomial time: Guess a harmful fact f and a subset of action occurrences O. Verify that f is true in the final state of the original plan π, but not in final state of the modified plan where O is removed from π.
- Hardness: *3SAT* can be reduced to *impermissibility*. Assume a 3SAT problem instance with n variables v_i and m clauses c_j. The planning instance has variables $V = \{v_1, \ldots, v_n, c_1, \ldots, c_m, b, g\}$, for each variable v_i an action $V_i : \langle \top, v_i \rangle$, for each clause $c_j = (l_{j1} \vee l_{j2} \vee l_{j3})$ an action $C_j : \langle \top, \bigwedge_{k=1}^{3} l_{jk} \rhd c_j \rangle$, the action $G : \langle \top, g \wedge (\bigwedge_{j=1}^{m} c_j) \rhd b \rangle$, and the action $B : \langle \top, \neg b \rangle$, with $u(\neg b) = -1$ and 0 for all others. Consider the plan $V_1, \ldots, V_n, C_1, \ldots, C_m, G, B$ on the empty initial state. If we can delete a subset of the V_i's so that the original formula becomes statisfiable then by deleting this set together with B, we show impermissibility. Similarly, impermissibility implies that the original formula is satisfiable.

Means to an end

Important notion: means to an end.

- When is an effect in a plan a means to an end?
- Use *counterfactual analysis*: Would the final intended (end) effect occur if the potential (means) effect did not happen?
- Light candle to make something visible.
- Switch light on and light candle . . . What is the means?
- Use toggle switches . . .

→ An effect in a plan is a means to an intended end effect, if this *end effect* were not true in the final state if *some subset* of the particular means effect is *deleted* in the plan.

Do-no-instrumental-harm plan validation

- The *means to an end* definition implies that we have the same combinatorial problem as for the simpler *do-no-harm principle*.

Theorem

The do-no-instrumental-harm plan validation problem is co-NP-complete.

Double-effect plan validation

- All criteria except for the *no negative consequence may be a means to the goal* condition can be checked easily.

Theorem

The do-no-instrumental-harm plan validation problem is co-NP-complete.

Complexity Summary

Ethical principle	Computational complexity
Deontology	linear time
Utilitarianism	PSPACE-complete
Do-no-harm principle	co-NP-complete
Asimovian principle	PSPACE-complete
Do-no-instrumental-harm principle	co-NP-complete
Doctrine of double effect	co-NP-complete

Summary

- Thinking about ethics in AI is unavoidable these days!
- There exist a number of ethical principles/guidelines from different institutions, which are very similar, though.
- In particular, fairness, privacy, and explainability have sparked new research directions in AI.
- Machine ethics is the field of covering ethics from a computational point of view.
- Self-driving cars have to cope with dilemma situations!
- There is no theory about ethics in action planning.
- Generalization of action-based to plan-based ethical judgments is possible.
- Surprising complexity results, based on the fact that the same effect can be made true arbitrarily often and can interact with each other.