Dina Babushkina
University of Twente, Jan. 2021

Moral Decision Making and AI: an Introduction

What is moral decision making?

Decision making – a procedure of arriving at a choice with the intent to act.

How to arrive at a choice? – reasons behind the choice

- actual reasons people use to choose;
- reasons that explain why A has done X

Descriptive level, explanatory reasons

The reasons that people should use to make right choices

Normative reasons

Ecomonic reasoning
(how to maximaze profit)

Moral reasoning
(how to do what is morally right/good)

…

Moral decision making - choosing a course of action based on moral reasons

Such decision:

- Must have a justification/explanation (we must be able to reconstruct the reasons)
  >> transparency, explainability and interpretability

- Must be based on moral reasons:
  be universalizable
  be impartial
  be objective
  appeal to intrinsic & incommensurable values

- Comes with responsibility

"Big no-no(s)" in moral decision making

"Ban on mixed worlds"

There can be no difference in moral properties of two phenomena without the difference in thier natural properties.

>> If it is right for Mike to X, then it cannot be that it is wrong for Sara to X
(if Mike and Sara are in the same situations)

No ought from is

One cannot derive norms from fact.

>> problem for "crowdsourcing moral views" as a strategy for creating ethical algorithms.

Why AI in relevant?

AI – a computational system that is interacting with the environment and is capable of directing this interaction without direct human control.

Moral decision by AI = decisions what would have morally significant impact (potential to cause harm) if they were made by a human.

Problem:

- Autonomy of AI (= independence from human control) in decisions that affect human well-being
- Black box instead of a decision making process
- Expectation for AI to act according to ethical principles & moral values
- Scale and the cost of mistakes

Areas of application:

- Ethical advisors
  - general
  - in medical sphere
  - moral prompters, e.g. Humane Eating Project
    http://www.americaforanimals.org/humane-eating-project/
  - etc

- Algorithms that make morally significant choices without human control:
  Autonomous military robots,
  Recruitment algorithms,
  Health applications,
  Various medical algorithms, incl. those deciding on the priority of care
  Autonomous vehicles etc.

How to program ethics to AI?

1. Top-down:
Set the criterion of rightness → let the algorithm apply it to a specific situation

2. Bottom-up:
- Let the machine develop into a moral agent
- Let the machine generalize its own moral rules

3. Hybrid approaches:
Combinations of top-down and bottom-up

Top-down approaches: the main ethical theories

Consequentialist approaches

These focus on the ability of a system to apply the consequentialist principle in one of its forms (there are some variations in the theory) to a specific situation.

Basic idea: CONSEQUENCES

the action is right if and only if it
brings about the best possible consequences for all

In combination with a belief that the only thing that is good in itself is utility/pleasure ->
the action is right if and only if it
maximizes utility/pleasure

See more : https://youtu.be/51DZteag74A

Problems:

- Intention does not matter
- Does not exclude harm/pain: ends justify means
- Does not care about individuals
- Close to prudential thinking borrowed from economic world

- Utility/please are too abstract–open to interpretation
- Extent and relevance of consequences

- Computationally challenging: To be fair in consequentialist terms, each decision would have to be based on the complete knowledge (all alternative, all interested parties, etc.). It is unclear how to achieve it + requires a lot of computational power.

Deontology – rightness/wrongness of actions themselves

Basic idea: RULES

The action is right if and only if it is prescribed by a rule.

E.g.:
- Ten Commandments
- the golden rule : "Treat others as you would have them to treat you"
- the ten commandments for C Programmers, Lysator, the Academic Computer Society, Linköping University in Linköping, Sweden https://www.lysator.liu.se/c/ten-commandments.html

Problems:
- Are not universalizable
- Often apply only to a certain group of people
- Usually justified with the reference to authority or convention
- Often against intuition

See more: https://ethics.org.au/ethics-explainer-deontology/

Deontology (Kantianism)

Basic idea: THE UNIVERSAL LAW

The action is right if and only if it is universalizable, i.e. can be required from anyone in the same situation.

- Morally right action is an expression of respect to humanity in every person (equal importance and rationality)
- Rules that we all can rationally consent to
- What is morally right does not depend on what we want (categorical)

See more: https://youtu.be/2S_XuJTOEJY

Problems:

- Does not look at the consequences of actions
- Does not include non-rational beings
- Computationally challenging: the system will have to be able to reason whether an action that it is about to execute satisfies the universalizability requirement. It is hard to see how one could create such an algorithm.

Virtue ethics

Basic idea: A WISE, VIRTUOUS PERSON

The action is right if and only if it is something a virtuous person would do.

See more https://youtu.be/NMblKpkKYao

Problems:

- No actions is wrong in itself
- Hard to define what a virtuous person is
- Hard to apply in daily practice, but maybe suitable for such context as relationships & some professional codes of conduct
- Risk of paternalism
- Computationally challenging: even if we agreed upon moral virtues, these are complex patterns of mental states (such as beliefs and desires), how could one create such patterns in a machine?

Examples:

MoralDM combines deontology with consequentialism

Jeremy is a consequentialist algorithm.

Bottom-up

Create artificial environments in which an AI system will be able to generate ethically acceptable solutions without any pre-programmed ethical principles. Two options.

(1) an expectation of an emergence of an ethical agent. Moral behaviour and moral reasoning are expected to emerge as a result of a certain level of complexity of non-moral capacities of a machine, such as self-interest, curiosity, planning etc. They rely on the analogy with children who are not born moral agents but learn to become moral agent by mimicking others. The expectation is that an AI system will first achieve a child-like moral maturity and then gradually evolve into an adult agent.

- Not possible at the moment, if at all.
- If possible, how to guarantee that the AA will be any better at moral judgement that any random human.

(2) Training an AI system on a data set of cases with ethically acceptable solutions to derive some sort of ethical rules or principles of moral decision making.

- Interesting and promising, but
- Risks to reproduce bad moral reasoning of humans,
- Risk of bias,
- Risk of deriving ought from is,
- Problem with applicability (difference between training and prediction datasets).

Examples: GenEth, Genetic Algorithms

Hybrid approaches

This cluster of solutions wants to overcome the limitations of top-down and bottom-up approaches and targets a combinatory approach to the problem of moral decision making by AI systems. The idea here is neither rules alone, nor the bottom-up strategies are enough for the moral decision making.

Examples: LIDA, Oracle AI, and MOOD.

Thank you!

Contact me: d.babushkina@utwente.nl

Title picture is by
Ramdlon at pixabay