Workshop on Ethics in the Design of Intelligent Agents (EDIA'16) at ECAI'2016

The Hague, August 30th 2016

A brief report

35 people had registered to EDIA and an average of 25 were present in the room all day long. Bertram Malle's and Jeremy Pitt's invited talks together with the contributed papers authors' presentations generated many questions and heated debates.

Two main topics were discussed: i) Values in the design of intelligent applications and ii) Formal models for ethical decision making. As papers can be found in the CEUR proceedings, only the main points and questions are reported here.

I - Values in the design of intelligent applications

[Malle] Robots with special roles, tasks, context have to have specific moral competence. But if you want to implement values, you have to know what it is. Values considered here as an abstraction of norms. Norms are hierarchical, fast-activated, contextual, updated. Therefore a hierarchical system of moral concepts and terms is built thanks to the results of experiments with participants. In the experiments, participants face a specific context picture (a library, a restaurant, etc.) and are asked to cite what is prescribed, permitted and prohibited in such contexts. It appears that permissions are more cited than prohibitions (e.g. "don't kill does not appear).The context itself is a hierarchical network. Context may be physical, temporal, linked with roles, goals, etc. Therefore the robot has to be a good perceiving system.

If you don't know a context, you narrow your norm network down (number of norms, etc.) You cannot take one-to-one human and robot. Nevertheless the properties that are found with experiments involving participants can constrain the computation. *Recommendation: you shouldn't build the universal robot!*

[Mermet] Ethical rules give properties to moral rules in a given context, e.g. they can order, or change priorities on moral rules according to the context

[Verdiesen] How to differentiate facts from values? What is a fact? What is a value? Truth is subjective. Which values have I? *Exercise: record my values for two months.*

[Cyra] Can virtual assistants be legitimized to request and provide meaningful reminders? *Recommendation: you shouldn't copy what people do, but design practical things.*

[Dogan] Practical case of autonomous car. Case studies: time and space decision (left turn, stuck behind roundabouts), extensions of trolley dilemma. Four moral profiles are considered: egoist (I), risk adverse altruist (You and I), sufficient altruist (We), pure altruist (You).

[Theodorou] The machine nature of a robot should be transparent: decision-making mechanism should be exposed, i.e. deal with relevance, abstraction, presentation of information. *Recommendation : transparency should match the level of understanding / interest of the user. Warning: human nature tends to assign moral agency to animals and objects.*

[Pitt] What does it mean, or require, to be responsible designers of intelligent agents, systems, or social-technical systems? Some pathologies: metrication of values, commodification of values (social values are replaced with a purely monetary value), dissolution of values, neo-colonialism (work at the edge, profit in the middleware, taxation nowhere). *Recommendation: need of some kind of Hippocratic oath for computer scientists.*

II - Formal models for ethical decision making

[Rolf] What if robots create new (i.e. not pre-programmed) goals? How to make the robot ethical? Robots might have to learn new ethical rules.

[Bringsjord] Logic frameworks may fail to represent ethically correct intelligent agents (e.g. first order logic is slow, naïve encoding may be unsound, etc.) Indeed virtues ethics is based on a rejection of obligations, prohibitions, norms, etc!

[Rocha Costa] Law is a centralized (normative) system whereas moral is a decentralized system. In a social system, goals may be inconsistent and several moral systems may exist and conflict. Moral norms, moral facts (actions and judgements) and moral judgements should be distinguished. A moral judgement is not a regulation without a motivation to be judged in a good way. So does it reduce to reward and punishment?

[Bonnemains] Judging ethical dilemma situations from the point of view of different ethical frameworks: this is not a human point of view, so what to do with the results? Is there a meta-ethical framework to choose what to do? What does it mean to justify a choice?

[Cointe] Ethical asset management: how to deal with long term consequences (e.g. agent avoid unethical assets, therefore financial value of assets goes down, and people buy them). How to model and take into account others ethics and judgements?

And yet more controversy about the Trolley problem:

- the trolley situation will never happen: indeed agents are connected (traffic control devices, other connected cars) and data come from all devices. Lin's point of view is naïve.
- it is not a real thing (only two options, no uncertainty).
- but it allows to push people in different directions and see divergences of ethics clearly.