The Future of AI: Ethical, Legal, and Societal Issues", 28th January – 1st February

28 January – 01 February 2019 @Snellius

The goal of the workshop was to create a multidisciplinary research agenda to bridge the conceptual gap among disciplines involved in developing AI applications, and to raise the level of sophistication at which designers reason about the impact of these applications. The led us to the following four questions, to be discussed during the week.

1. ***What are the legal or ethical impacts of AI systems?***
2. ***Can we define a unified conceptual framework to reason about these impacts, usable across all disciplines involved in designing AI systems?***
3. ***Can we build legally-aware or ethically-aware agents that use such a framework?***
4. ***What are the legal or ethical impacts of AI systems that themselves make legal or ethical decisions?***

We discussed each question during one day and devoted the 5th day to drawing conclusions from the summaries of the discussions, and to a panel discussion open to the public.

1. What are the legal or ethical impacts of AI systems?

The impact most frequently discussed was the diffusion of responsibility if AI systems make decisions. The decision context typically consists of a network of people and machines and it may not be clear who has the final responsibility of a decision. This is called the responsibility gap. If we declare final responsibility always to be with a person, then by definition, the responsibility gap has been resolved. But this is difficult to do in the case of technological action. One could also say that there is a problem of responsibility attribution. And to the extent that we place responsibility with machines, people's ability to take responsibility may disappear, which may be called ethical deskilling.

Several solutions were discussed, such posting ethical warnings by AI and relating ethical decisions to the specifics of a context. Most participants agreed that responsibility needs to remain with the human.

2. Can we define a unified conceptual framework to reason about these impacts, usable across all disciplines involved in designing AI systems?

Creating a unified conceptual framework that crosses disciplines is possible, and we made first steps towards such a framework.

- **Autonomy** - the perceived and experienced ability to set own goals - e.g. the decisions the system is making for you would not be one of your own; autonomous cars increase autonomy by relieving people of driving; Alexa making a call and recording the conversations of people, etc.
- **Agency** - the perceived and experienced ability to act - e.g. over/under attributing agency to AI systems
- **Misrepresentation/misconception** - (intentionally) framing/perceiving an AI-based product in a specific way that overestimates/underscores its human-like capabilities (the title itself - intelligence)
- **Anthropomorphism** - attributing human-like properties to technologies - e.g. Alexa as a normative guideline

- **Interaction** - the way AI-based product influences the human relations to the self, others and the sociotechnical world - e.g. AI whistleblower - internalizing surveillance, decrease in trust, delegation of civic duties to AI...
- **Fairness** - treatment of people based on as objective criteria as possible / equally taking into account the interests of all stakeholders - e.g. Alexa not understanding accents
- **Responsibility** - attribution of blame/praise in the context where people interact with AI - e.g. the self-driving car killing someone
- **Safety** - ensuring the physical and mental conditions of individuals/the optimal performance of software within the preset bounds - e.g. the direction of learning, lack of control, inefficient safety assessment of self-driving systems, training of cars on rural areas and implementing in urban settings
- **Security** - resilience against malicious actors - e.g. hacking of/with the AI systems
- **Privacy** - proportional and balanced management of individual's data - e.g. data misappropriation
- **Control** - an ability to monitor and intervene into the algorithm performance - e.g. transition of control - the car giving back control to you when you are not prepared and didn't expect it
- **Trust** - fulfillment of positive expectations regarding the AI products - e.g. AI assistants can't have a human conversation
- **Robustness** - correct functioning in unpredictable environmental
- **Predictability** - accurate anticipation of future behavior
- **Transparency** - the ability at any point to provide an understandable and meaningful explanation on a status of a system

To make a framework out of this, relations among concepts need to be defined, and the relation with definitions in various fields must be explicated. In addition, to test the framework the response of the leaders of the various disciplines should be collected and the practical usability tested.

3. Can we build legally-aware or ethically-aware agents that use such a framework?

We can build ethically- and legally-aware agents, but if they can use a framework such as the one outlined above remains to be seen. AI systems make assumptions about their context that may turn out to be unrealistic in particular cases. For example, a driver of an autonomous car can easily game the system into thinking that the driver has his or her hands on the wheel, by tying a bottle to the steering wheel. This raised the question whether drivers of autonomous cars should get a special driver’s license that certifies certain skills, just as operators of nuclear plants need special training to monitor the autonomous systems that run the plant. Where the autonomous system fails due to incorrect assumptions about its context, the human operator must step in to take action. This takes us back to the question who is responsible for failure of the human-machine system.

More fundamentally, machines as we know them are not conscious, and from the beginning in the Darthmouth workshop, AI has always been defined as an imitation of intelligence. Hence so-called “ethical” agents are also merely imitating. This raises questions whether the term “aware” should be used at all, and whether an imitation of ethical awareness is sufficient.

4. What are the legal or ethical impacts of AI systems that themselves make legal or ethical decisions?

The legal or ethical impacts of AI systems that make legal or ethical decisions are as yet unexplored and require adaptation of law. Insurance contracts need to be adapted too, to determine who pays the premium. Distributing the insurance premium over participants in a situation, including the manufacturers of an AI, implies how responsibility is distributed in an otherwise diffuse situation. There was also some discussion during the workshop about whether or not AI systems can take "decisions" at all.

The conclusion of the workshop participants for a research program is that we need to somehow counterbalance AI as big business, where powerful companies from the U.S.A. and China dominate the development and application of AI systems. The goals of these companies do not necessarily harmonize with ethical and legal goals in Europe. This brought up the suggestion that we should develop business models for ethical AI, and should stimulate AI by small companies for small companies. A relevant goal for AI research should be to empower people and to protect them from the adverse impact of AI on their personal life.

We ended the workshop in the morning of day 5 with a first list of relevant research questions, to be refined, elaborated and structured in later discussions:

- How can we understand and govern AI-enabled commons?
- What are the governance needs of large-scale sociotechnical AI systems?
- Which regulations are needed to protect people working with AI systems in an organization?
- What do we need to measure AI impact?
- Define our multidisciplinary conceptual Framework?
- Can we build a falsification / adversary machine?
- How do we / should we identify value concerns in the design process of AI systems?
- How does an artifact imbue values?
- What values does a given system incorporate?
- How do you design a socio-technical system that induces certain values?
- What legal framework do we need for AI systems design?
- How can we embed AI systems in an organization such that human on the loop don't suffer from the inversion of the burden of proof?
- Develop empirical and theoretical research methods to test human-AI systems.
- What vulnerabilities are introduced by AI and how can they be mitigated?
- How can we theorize context in AI design?
- How can we empower users to survive ethical trespassing?

The panel members during the public discussion of the afternoon of day 5 were

- Maxim Februari (writer, philosopher of law, columnist NRC)
- Benny Mols (free lance science journalist, specialised in artificial intelligence and robotics)
- Cathelijn Muller (President ALLAI & member of the EU High Level Expert Group on AI)
- Alexander Rinnooy Kan (member of the senate, UvA)
- Peter Paul Verbeek (Philosophy Lab, Un. Of Twente)
- Aimee van Wijnsberghe (TU Delft and Foundation of Responsible Robotics, co-founder ALLAI)

Moderator was Virginia Dignum. Some of the key issues discussed were the global applicability of AI systems, which makes them relevant for all of us and calls for global guidelines for ethical design and use of AI. At the same time, these guideline should be usable for decision-makers in practice. One of the most intriguing questions asked from the public is the question of fairness. As a matter of fact data sets used in AI are biased and we know that this is unavoidable. Statistically, bias evens out if we take averages over a long run of random sampling from the same population. But in AI we draw one sample, and this sample will contain bias. Who determines whether this data set is fair enough to be used? How is this determined? Who is accountable for this decision? How do we harmonize different views of fairness, and how do we deal with changing perceptions of fairness?

A more detailed report on this panel is available at http://allai.nl/highlights-of-the-public-closing-event-of-lorentz-workshop-the-future-of-ai/.

Virginia Dignum (Umeå, Sweden)
Roel Wieringa (Twente, Netherlands)
Mark Coeckelberg (Vienna, Austria)
Ugo Pagallo (Turin, Italy)