Dutch Artificial Intelligence Manifesto

Artificial Intelligence is the science and engineering of intelligent systems. AI systems are capable of sensing their environment, learn from and reason about it, and change it based on advanced decision making. AI already has had a big impact on our society but this impact will only increase due to recent developments and ongoing trends in machine learning and robotics. AI applications will become ubiquitous in all areas of our society including industry, transportation, healthcare, education, and public safety. As a consequence, AI is transforming our society by changing the nature of work and the way humans relate to machines. The key to using Artificial Intelligence effectively, however, is to augment and assist humans, not to replace them. Figuring out how to get humans and AI to collaborate effectively will have big economic impact, and will shape the way the workforce is educated. The challenge therefore is how to re-define the relationship between humans and machines that will soon be able to perform advanced tasks. While delegating such tasks to AI systems we will also need to engineer reliable and robust AI systems to ensure acceptance and trust of their users. Addressing this challenge requires a fundamental re-orientation of AI that takes their human users seriously. We argue that the following key questions need to be investigated:

- AI systems will become proactive and personalised and will (need to) collaborate with humans. How will this redefine the social interface between humans and AI systems?
- AI systems will perform sophisticated and advanced tasks that were performed by humans and moreover may significantly affect their human users by performing these tasks. How can we make sure that we will understand why an AI system affected us the way it did?
- AI systems enable the automation of tasks that humans used to perform and the automated processing of huge amounts of data. How can we ensure that the results of those actions and of data processing match our expectations and moral standards?

In order to address these questions, we need to focus on three areas in AI research for the coming decade:

- **Socially-Aware AI**: AI that is able to interpret, reason about, and influence human behaviour, resulting in intelligent systems that interact, collaborate and coordinate their behaviour with human beings.
- **Explainable AI**: AI that is able to explain to its users in an intuitive, human-understandable manner how and why its underlying algorithms produced the AI's behaviour.
- **Responsible AI**: AI that is able to take into account moral, societal and legal values, while efficiently processing the abundance of (sensory) information available worldwide.

Socially-Aware AI

Technological developments over the past decades are reshaping artificial intelligence in various ways, both concerning environments in which intelligent systems are being deployed and concerning environments in which data is being collected. In particular, modern AI systems should be able to understand and reason about their social context, allowing them to interact and collaborate with human beings in order to achieve joint goals. Examples of systems that require this 'social' ability include smart environments (cities, buildings, rooms), social robots, wearable devices (e.g., for health tracking) and a range of handheld devices. Socially aware AI aims to leverage large-scale,

dynamic, continuous, and real-time sensory data as well as computational models of physiological and cognitive processes to recognize individual behavior, discover group interaction patterns, and support human collaboration.

Explainable AI

As AI systems such as robots and machine learning and decision making algorithms will significantly affect their users, it is important to be able to explain how and why an AI system produced the effect that it did. In many practical cases (e.g., healthcare, security) it matters how the results of a decision came about and due to the increased complexity of AI systems some form of explanation from such systems will be required. It is a well-known hurdle for today's data-driven artificial intelligence that the algorithms produced behave largely as black boxes. For instance, algorithms trained by extensive data analysis using state-of-the-art deep learning techniques perform well in terms of the input-output function they represent, but it remains hard to make sense of the internal structure of the learnt algorithms. In order to make progress, we believe that there is a need for investing in explainable AI, i.e., in data analysis and machine learning techniques that develop understanding of the available data, e.g., in the form of complex knowledge structures. The aim is to design tools to help make the inner workings of AI systems more transparent.

Responsible AI

Today’s complex AI systems have become less predictable while at the same time their potential impact on our society has increased. In order to be able to accept and trust these systems we will need tools and techniques to make sure these systems will comply with our norms and standards. Data-driven artificial intelligence techniques are designed for, hence good at discovering patterns in the data. Because of the aim of objective data analysis, both wanted and unwanted patterns are discovered. For instance, machine learning algorithms run the risk of adopting the same racial, gender or other biases that humans have (e.g., ‘most nurses are female’). Even when such patterns are present in the available data, the consequences may have critical consequences for society. What is needed is data analysis technology that can be steered normatively, so that the outcomes meet ethical criteria. This requires the balancing of objective, descriptive aims and subjective, normative aims (cf. what is common in social sciences such as the law). Great impact can be expected from such ethical systems design methods, also with the perspective of ever more autonomous weaponry and privacy-invading investigative measures. New frameworks are needed that can guide us in identifying the moral issues that arise due to the application of AI and help us determine the parameters that we as a society want to be optimized in these systems. We need techniques that support verifying that the behaviour of an AI system stays within those parameters. These techniques may also provide an alternative when an AI system is not capable of yielding explanations. Moreover, responsible AI also includes the challenge to process the abundance of available data in an efficient and environmentally friendly manner.