Responsible Artificial Intelligence: Ethical Thinking by and about AI

Virginia Dignum, Delft University of Technology, The Netherlands, m.v.dignum@tudelft.nl

As advances in AI occur at high speed, many questions raise across social, economic, political, technological, legal, ethical and philosophical issues. Can machines make moral decisions? Should artificial systems ever be treated as ethical entities? What are the legal and ethical consequences of human enhancement technologies, or cyber-genetic technologies? What are the consequences of extended government, corporate, and other organisational access to knowledge and predictions concerning citizen behaviour? How can moral, societal and legal values be part of the design process? How and when should governments and the general public intervene?

Answering these and related questions requires a whole new understanding of Ethics with respect to control and autonomy, in the changing socio-technical reality. The urgency of these issues is acknowledged by researchers and policy makers alike. Moreover, implementing ethical actions in machines will help us better understand ethics overall. To enable the required technological developments and responses, AI researchers and practitioners will need to be able to take moral, societal and legal values into account in the design of AI systems. AI researchers are designers for values, who can elicit and represent human values, translate these values into technical requirements, who can innovate in cases of moral overload when numerous values are to be incorporated, and who can demonstrate that design solutions realize the values wished for.

RAI: Who is responsible?

In our view, there are three aspects of particular concern, if we want to ensure that AI-related developments are to be for societal good. Firstly, awareness of the ethical consequences of AI research and development, and formalisation of accountability. Who is to blame if a self-driving car harms a pedestrian? The builder of the hardware (sensors, actuators)? The builder of the software that enables the car to decide on a path? The authorities that allow the car in the road? The owner that can personalise the car decision-making system to meet its preferences? The car itself because its behaviour is based on its own learning? All of them?

Secondly, the need to develop models and algorithms that enable AI systems to reason about and take decisions based on responsibility, and to justify their decisions accordingly. Current deep-learning mechanisms are unable to link decisions to inputs, and therefore not explain their acts in ways that we can understand. However, we need our future self-driving cars to deal with moral dilemmas such as the decision to veer left and harm a pedestrian or veer right and harm its passengers.

Thirdly, transparency and participation is necessary. Here education plays an important role, both to ensure that knowledge of the potential AI is widespread, as well as to make people aware that they can participate in shaping the societal development. A new and more ambitious form of governance is one of the most pressing needs in order to ensure that inevitable AI advances will serve societal good. Only then accountability, responsibility and transparency are possible.

The ethical implications of AI include liability and law. For example, who is liable if a driverless car is involved in an accident? Should AI be covered by existing cyber law, or should it have specific rules? What rules should be made to control the deployment of autonomous weapons? That is, means are needed to integrate moral, societal and legal values with technological developments in AI. Responsible AI is more than the ticking of

some ethical ‘boxes’ or the development of some add-on features in AI systems. Rather, responsibility is fundamental to intelligence and to AI research. For example, understanding how Deep Learning-based advances in computer vision should go hand in hand with ethical considerations on their use as autonomous “decision makers” in target identification.

However, Responsible AI is not just about making rules to govern intelligent machines – we also need to consider how we regulate the data they create and share. It is time to make up our minds and respond to the novel, data-driven reality with a constructive operational framework that is able to inform sustainable new modes of political thinking. Digital technologies and data science are now used to shape our societies, to constitute the very fabric of our sociality, often circumventing democratic decision-making and scientific facts.
Regulators and device manufacturers need to consider that connected devices provide extra opportunities for both legitimate organisations and hackers to access personal data. The statement "code is law" reflects the fact that algorithms increasingly shape our reality.

While the behaviour of people is regulated by hundreds of laws, algorithms are subject to very few regulations, even though they may have super-human powers. This is inappropriate and dangerous, as actors without legitimate grounds increasingly interfere with our lives, often without our knowledge. One should not allow this, even for a free service. Humans are increasingly losing control over the information and communication technologies they have created. Currently, cybercrime costs us 3 trillion dollars annually, and it is increasing exponentially. Therefore, it is crucial to create the tools and institutions that support the transparent design of technological systems that are compatible with our moral, social and cultural values such as (informational) self-determination, security, sustainability, democracy, participation, safety, transparency, accountability, and the emergence of certain properties and functions. A system that maximizes single indicators (such as citizen scores or gross national product per capita) does not serve humans well. It is essential to accommodate value pluralism and learn, for example, to design for efficiency, usability, flexibility, resilience, fairness, justice, dignity, happiness, well-being, safety, security, health, empathy, friendship, solidarity, and peace.

AI for labor and welfare

Currently, over a million persons die annually in traffic accidents, more than half of which are caused by human error. Even if intelligent self-driving cars will inevitably cause accidents and deaths, forecasts show a sharp decrease on road casualties associated to increase in self-driving cars. Similarly, jobs will be lost, but maybe repetitive, monotonous, demeaning jobs should be lost, freeing people to more meaningful and joyful occupation. AI developments will contribute to a needed redefinition of fundamental human values, including our current understanding of work, wealth and responsibility.

Work. As AI systems replace people in many traditional jobs, we must rethink the meaning of *work*. Jobs change but more importantly the character of jobs will change. Meaningful occupations are those that contribute to the welfare of society, the fulfilment of oneself and the advance of mankind. These are not necessarily equated with current ‘paid jobs’. AI systems can free us to, and be reward for, care for each other, engage in arts, hobbies and sports, enjoy nature, and, meditate, i.e. those things that give us energy and make us happy.
Wealth. Technological developments in the last century led to mass production and mass consumption. Until very recently, *having* is the main goal, and competition the main drive: “I am what I have”. Digital developments, including AI, favour openness over competition: Open data, open source, open access, … The drive is now quickly shifting to sharing: “I am what I share”. Combined with the changing role of work, this novel view on wealth, requires a new view on economy and finance.
Responsibility. As AI moves from a tool to teammate, perhaps the most important result of AI advances is the need to rethink *responsibility*. Developments in autonomy and machine

learning are rapidly enabling AI systems to decide and act without direct human control. Greater autonomy must come with greater responsibility, even when the notions of machine autonomy and responsibility are necessarily different than those that apply to people. Machines are already making decisions. We need to deal with longer chains of responsibility, and with responsibility being extended to refer to machines and corporations.

Responsibility contributes to *trust* and includes *accountability*, i.e. being able to explain and justify decisions. Whereas our trust on other people is partly based on our ability to understand their ways of doing (by putting ourselves on their place), this does not go for machines. Trust on machines must then be based on *transparency*. Algorithm development has so far been led by the goal of improving performance, leading to opaque black boxes. Putting human values at the core of AI systems calls for a mind-shift of researchers and developers towards the goal of improving transparency rather than performance, which will lead to novel and exciting algorithms. Turning Deep Learning into Valuable Learning.

RAI challenges

The current raise of AI is often compared to the Industrial Revolution. Just recently Steven Cave, director of the Leverhulme Centre for the Future of Intelligence, said in an interview with the Telegraph: "AI revolution is likely to happen even faster - so the potential damage is even greater." We, AI researchers and developers are the ones making this possible, it is up to us to do something about it!

In fact, machines have been making decisions for us for quite some time. My thermostat decides whether to turn the central heating on or off based on the information it has about my preferred room temperature, the electronic ports in my local train station decide whether to grant me access to the platforms based on the information it gets on my travel credit and on constraints given by the local travel authorities, Google decides which of the trillions of internet pages I am more likely to want to see when I search about Trump based on the information it has about my preferences (and will be more likely to show me pages critical of Trump than someone who voted for him). So, machine decisions are nothing new, what makes AI decision different, and frightening to many, is the opacity of its decisions, i.e. that AI decisions are often made without human input. We are the ones that determine the optimization goals and the utility functions at the basis of machine learning algorithms, we decide what the machine should be maximising. Indeed, even in the famous paperclip maximiser example by Nick Bostrom, someone gave once this paperclip maximisation goal to this unfortunate intelligent factory.

AI clearly has the potential to transform the way we live and work, but it is important that we set appropriate limitations and controls. Otherwise the risk is that rather than expanding our horizons and our potential, the compromises that we are already making in terms of access to our personal information could end up compromising our choices, and even our basic human rights.

Concluding remarks

AI potentially poses many risks to human rights, as e.g. recently described in the openDemocracy site. However, more than a risk for human values, AI brings in itself enormous potential to improve the lives of many, and to ensure human rights to all. A group of researchers and developers have joined forces on the IEEE's Global Initiative for Ethical Considerations in the Design of Autonomous Systems, of which I am a member of the executive committee, and just recently launched a document on Ethically Aligned Design highlighting the importance of empowering developers to prioritize ethical considerations in AI.

It is up to us to decide. Are we building algorithms to maximize shareholder profit or to maximise fair distribution of resources in a community, by providing solutions to tragedy of the commons situations and ensure free access to information and education to all; to optimise company performance or to optimize crop yield for small farmers around the world, by proving real-time information on fertilizer levels, planting and harvesting moments and weather conditions; or to improve proficiency in playing Go or to improve cross-cultural communication, by providing better contextualised translation services?

It is up to us to decide. Are we following the money or following the society best interests? Are we basing AI developments on shareholder value or on human rights and human values?

It is up to us to decide. Are we standing on the side-lines twitting, blogging and writing opinion articles on the potential dangers of AI, or are we taking action to ensure AI development centred on human values, benefiting all mankind?

Responsible AI. We are responsible.