RESPONSIBLE ARTIFICIAL INTELLIGENCE: DESIGNING AI FOR HUMAN VALUES

Virginia Dignum
Delft University of Technology, The Netherlands

Abstract – Artificial intelligence (AI) is increasingly affecting our lives in smaller or greater ways. In order to ensure that systems will uphold human values, design methods are needed that incorporate ethical principles and address societal concerns. In this paper, we explore the impact of AI in the case of the expected effects on the European labor market, and propose the accountability, responsibility and transparency (ART) design principles for the development of AI systems that are sensitive to human values.

Keywords – Artificial intelligence, design for values, ethics, societal impact

1. INTRODUCTION

Artificial intelligence (AI) is becoming rapidly present in all aspects of everyday life. It is everywhere, it affects everyone, and its capabilities are evolving extremely rapidly. AI can help us in many ways: it can perform hard, dangerous or boring work for us; it can help us to save lives and cope with disasters; and, it can entertain us and make our daily life more comfortable. AI systems manage complex, data-intensive tasks, e.g. monitoring credit card systems for fraudulent behavior, enabling high-frequency stock trading, supporting medical diagnoses and detecting cybersecurity threats. Embodied as robots, AI is soon to move and work among us, in the form of service, transportation, medical and military robots. Nevertheless, current perceptions and expectations regarding the capabilities of AI vary widely and consensus on the societal impact of AI is hard to find. In the first part of this paper, we analyze this situation by means of a study on the expected effect of AI on the European job market.

The second part of the paper explores the social, economic, political, technological, legal, ethical and philosophical questions raised by AI and how design methods can deal with these. Currently, there is an increasing awareness that a responsible approach to AI is needed to ensure the safe, beneficial and fair use of AI technologies. This also includes the need to consider the ethical implications of decisions made by machines, and to define the legal status of AI. However, concrete approaches to the responsible design of AI are mostly non-existent. The responsible design, development and use of AI systems is of the utmost relevance to AI applications such as self-driving vehicles, companion, healthcare robots, and ranking and profiling algorithms, which are already affecting society or will be in a few years. In all these applications, AI reasoning should be able to take into account societal values, moral and ethical considerations, weigh up the respective priorities of values held by stakeholders and in different multicultural contexts, explain its reasoning and guarantee transparency.

Answering these and related questions requires a whole new understanding of ethics and to rethink the concept of agency in the changing socio-technical reality. 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. Developing AI responsibly requires the means to elicit and represent human values, translate these values into technical requirements, develop the means to deal with moral dilemmas and values preferences, and to evaluate systems in terms of their contribution to human wellbeing.

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 these notions are necessarily different when applied to machines than to people. Ensuring that systems are designed responsibly contributes to our trust of their behavior, and requires both accountability, i.e. being able to explain and justify decisions, and transparency, i.e. understanding the ways systems make decisions and how the data is being used, collected and governed. To this effect, we have proposed the principles of accountability, responsibility and transparency (ART) [7]. ART implements a design for values approach [26, 10], to ensure that human values and ethical principles, and their priorities and choices are explicitly included in the design processes in a transparent and systematic manner.

2. EXPECTATIONS ON THE IMPACT OF AI

In the past technical innovation has always created more jobs and led to a higher average standard of living; however, this does not mean that the implementation of new technologies has ever gone without opposition [4]. As shown by the Luddite movement in the 18th century and superbly demonstrated in Charlie Chaplin's influential movie "Modern Times", technological change and the subsequent displacement or change in the nature of jobs has led to great social unrest in the past [27].

The current wave of AI development has already incited wide public discussion on its effects on jobs and standards of living. An increasing number of people and organizations are warning about the possible negative impact of AI implementation on jobs and society, and several expect AI to cause more extreme effects than previous technological revolutions [4].

Boasting one of the world's largest economies and a highly educated workforce, this problem is very relevant to the European Union. Already in 2014, European Commissioner Kroes indicated that up to 70% of EU citizens believe that robots will steal people's jobs"[1].

Somber predictions on future AI capabilities put an increasing pressure on policy makers to protect the European economy and workforce. However, comparing possible policies proves to be hard given the uncertainty of future effects of AI. In fact, current studies on the influence of AI on the jobs market vary from a Utopian society in which nobody has to work, to the ending of economic growth in the western world [13, 23].

In order to provide European policy makers with a clear forecast of the future effects of AI on the European labor market and a recommendation on future policy directions combating potential harmful effects to this market we have performed a qualitative study on the expectations on AI. This forecast will be constructed by means of an adapted Delphi method study, facilitating discussion among European AI experts to create a consensus-based forecast of future AI effects on the European labor market.

2.1 Literature analysis

Existing reports on the number of jobs that can theoretically be replaced by AI in the long term, indicate figures as high as 47% of job losses in the US [9], and 35% in the UK [6]. Nevertheless, policy discussions on the effects of AI on jobs are still scarce in Europe. The topic seems to be of importance to some national governments [27, 6] but there is no clear European policy vision on potential harmful effects on the jobs market. In other countries, namely the USA, protecting workers from technological change is a more regular policy topic [20]. Moreover, few studies have provided a clear estimate of the amount of jobs that will be created or on the nature of future jobs. Some researchers looking at historical data expect that created jobs will outnumber those lost [18]. On the other hand, [25] states: "Experts envision automation and intelligent digital agents permeating vast areas of our work […], but they are divided on whether these advances will displace more jobs than they create".

Consensus does exist on the necessity of re-education of employees as preparation for future changes [9]. A panel of experts, hosted by McKinsey in 2014, expected that the number of US manufacturing jobs is rising and will continue to be in the coming years but it is very important to educate these people to work with machines otherwise they will not be needed in the future [16].

With regard to how AI contributes to this changing market, existing studies show an almost even divide among researchers between a positive and negative impact on the European economy [25]. Existing literature on this topic shows a very theoretical, sometimes philosophical, future view on labor markets. Testing these theories is hard, as they reflect the researchers' interpretation of existing data. Polling studies also show little consensus between researchers [25].

We use the four scenarios proposed by [27] as a means to classify the different studies:

A. **Business-as-usual:** According to this view, technological innovation always leads to higher productivity and the effect of AI will not be different. This productivity can in turn lead to either a larger or a smaller labor market, but, at

[1] http://europa.eu/rapid/press-release_SPEECH-14-421_en.htm

least in the long term, technological innovation has always had a positive effect on the number of European jobs [28]. The business-as-usual scenario therefore predicts a growth in the amount of European jobs market and economy in the long-term, coupled with a change in the nature of jobs and possibly short-term unrest. This unrest can be prevented with timely re-education of employees. Large wealth redistribution programs like the introduction of a basic income are not expected to be necessary.

B. **Techno-revolutionists:** According to this scenario, AI applications will in time compete with and take over an increasing number of human jobs. The deployment of autonomous systems will cause high levels of unemployment and create a growing gap between income from labor and income from assets, leading to an increasing divide in wealth. Major re-education policies are a necessity to make sure that humans will work with machines rather than compete with them for jobs, in a world where machines will outperform a majority of humans. Increasing wealth inequality is a result of big technological revolutions [21], leading to the need for a more balanced distribution of wealth [5]. This could lead to great societal challenges which require major (public) policy changes, such as the introduction of a basic income or a negative income tax. Tax incentives like subsidies for companies that keep humans on the payroll are also mentioned as policy options.

C. **Techno-utopists**: A small group of researchers expects that the exponential growth of technological developments will lead to negligible costs of information and energy, through which many physical goods and services will become (almost) free. Technological innovation, in this scenario, will eventually create a society of abundance rather than one of scarcity. Ownership and marginal costs will disappear, leading to the end of capitalism. According to [23], AI will be one of the enabling technologies for this scenario. As humans will spend less time on their jobs, and robots and computer programs will not have a salary, new forms of wealth distribution, such as a universal basic income will be needed to maintain the future economy [14].

D. **Techno-pessimists**: In contrary to the techno-revolutionists and the techno-utopists, techno-pessimists expect future economic growth to be lower than it is today. In fact [13] indicates that many innovations that can lead to strong economic growth are already implemented and cannot be repeated, whereas at the same time novel technological improvements fail to deliver strong economic effects. [13] Therefore expects future AI to have only a very limited impact on the European labor market. Techno-pessimists argue for increased policy to tackle existing economic headwinds rather than investment in AI.

Table 1 gives an overview of the expected impact of the different scenarios on the economic growth, the role of AI and the effect of different policies, based on the qualitative analysis of literature.

Table 1. An overview of the expected impact on economic growth and effects of different policy directions for the four scenarios: negative (-), neutral (0), positive (+) or very positive (++)

	A	**B**	**C**	**D**
Expected economic growth	Medium	Medium	High	High
Impact AI on economic growth	Medium	High	High	Low
Re-education	+	++	0	0
Wealth redistribution programs	+	+	++	0
Investment in AI	+	0	+	-
Subsidized human workforce	-	+	+	0

2.2 The views of AI experts

The scenarios presented in the previous clause highlight a fundamental disagreement on the impact of AI on the labor market and on the policies that are needed to regulate this impact. In this clause, we describe research performed at the Delft University of Technology in the Netherlands, using an adapted Delphi method [19] to facilitate an open discussion among researchers across Europe, This method combines the benefits of survey research, interview sessions and group discussions, and aims to identify the reasoning and rationale behind differences in opinions among AI experts while guarding against the occurrence of group think.

Delphi studies take an iterative approach to ensure that the strongest possible consensus among participants is reached by asking experts for their opinions on the combined results of previous rounds. The Delphi method does not state a fixed boundary on the amount of participants to form an adequate sample size. Finding motivated and knowledgeable respondents is more important than creating a statistically significant sample size. Delphi studies are often conducted with small sample sizes [22] and participants are not selected at random but because of their particular expertise. The most significant features of the Delphi method are its recursion and the possibility to get feedback and evaluate one's own answers. These characteristics of the Delphi method have been proven to guarantee the validity and reliability in case of studies aiming at predicting or understanding possible future scenarios [15, 11]. This method is therefore suitable to study the impact of AI on the European labor market.

In the above-mentioned study, that took place mid-2016, experts were invited to participate by email; emails were sent to relevant mailing lists and through the European Association for Artificial Intelligence (EurAI). All respondents were screened on their experience by the researchers. This approach led to a total of five respondents. Five additional experts, prominent European AI researchers from three different European countries, indicated their willingness to participate in an interview session for the validation of the study. The first questionnaire aimed at establishing an initial understanding of respondents' views on future AI capabilities, effects on the nature and size of the European labor markets, the factors influencing these effects and possible governmental roles and policies. The second questionnaire provided more detailed predictions through the identification of the timeline and specific factors expected to influence the effect of AI. As such, respondents were asked to comment on short to mid-term (0-10 years) and long-term (>10 years) effects. Respondents were further asked to reflect on specific policies, along the aspects identified in the literature study. The second questionnaire also included a section where respondents could rate the influence of a variety of factors on the effects AI will have on the amount of European jobs and on the nature of European jobs (on a scale of one to five). These results are depicted in Fig. 1.

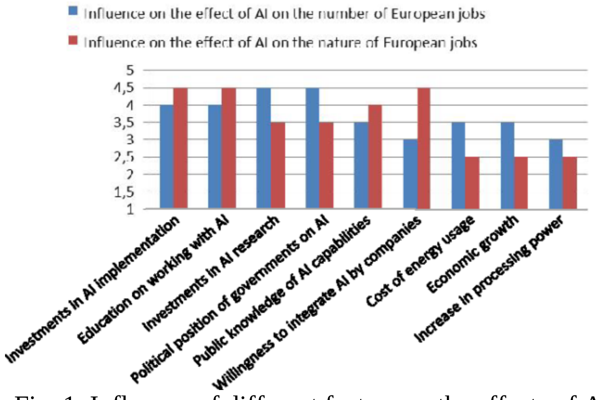

Fig. 1. Influence of different factors on the effects of AI on the nature and number of European jobs. Scores range from 1, no influence to 5, huge influence

After these two rounds of questionnaires, further validated by means of an interview session, respondents reached consensus on the following points: (i) Future AI will decrease the number of mechanical/non-knowledge intensive jobs in the short term; (ii) it will create new, most likely very specialized jobs; (iii) and will have a large impact on the nature of European jobs; (iv) governments will need to revise their education system to make sure their future workforce can work with AI. However, no consensus has been reached on the net result of the influence of AI on the number of European jobs in the long term and on the factors influencing the impact of AI on the number and nature of those jobs. In combination with the results of existing literature, described in the previous clause, this analysis of the views of leading AI researchers provides a useful forecast on the future effects of artificial intelligence on the European labor market to aid policy makers in preparing Europe for a smart future. The results of the Delphi study bring a somewhat moderate view on the effects of AI on the European labor market, which complement and extend current scientific literature. In their short term prediction, the views of the experts consulted are fairly consensual, and mostly aligned with the outcomes predicted by the business-as-usual scenario. Consensus on long term effects is narrower and includes elements from the business-as-usual, techno-optimist and techno-utopist scenarios. Nevertheless, it is important to note that Delphi style research leads to findings that are not necessarily statistically supported, but that can be used to inform further research on the expectations on the social impact of AI at a larger scale. The main contribution of this study is that it tempers the current hype on the impact of AI, by bringing in the views of AI experts with a long experience in the field. This can support policy makers in tempering their expectations.

3. RESPONSIBILITY IN AI

In this clause, we discuss how to approach the design of AI systems that are sensitive to moral principles and human value. 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 no system can be truly intelligent if it cannot understand responsibility.

Responsible AI rests in three pillars of equal importance. Firstly, society in general must be prepared to take responsibility for the impact of AI. This means that researchers and developers should be trained to be aware of their own responsibility where it concerns the development of AI systems with direct impact in society. This requires efforts in education and training and the development of codes of conduct. Moreover, responsible AI is an issue of regulation and legislation. It is up to governments and citizens to determine how issues of liability should be regulated. For example, who will be to blame if a self-driving car harms a pedestrian? Is it the builder of the hardware (e.g. of the sensors used by the car to perceive the environment)?; the builder of the software that enables the car to decide on a path?; the authorities that allow the car on the road?; the owner that personalized the car decision-making settings to meet her preferences?; or, the car itself because its behavior is based on its own learning? All these, and more questions must be informing the regulations that societies put in place towards responsible use of AI systems.

Secondly, responsible AI implies the need for mechanisms that enable AI systems themselves to reason about, and act according to, ethics and human values. This requires models and algorithms to represent and reason about, and take decisions based on, human values, and to justify their decisions according to their effect on those values. Current (deep-learning) mechanisms are unable to meaningfully link decisions to inputs, and therefore cannot explain their acts in ways that we can understand.

Thirdly, participation; it is necessary to understand how different people work with and live with AI technologies across cultures in order to develop frameworks for responsible AI. In fact, AI does not stand in itself, but must be understood as part of socio-technical relations. Here again 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.

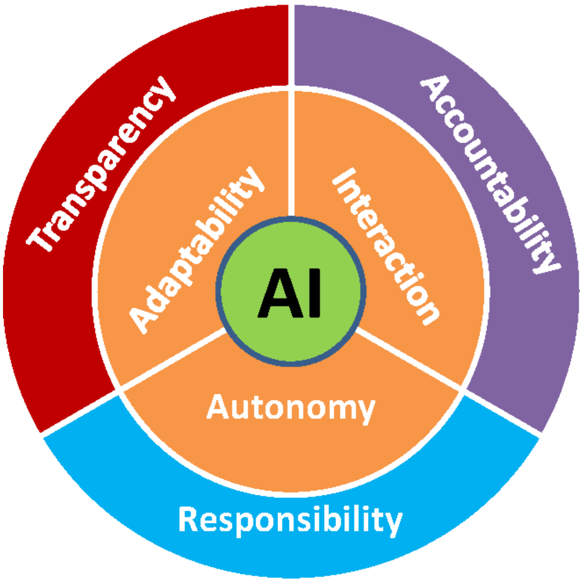

Fig. 2: The ART principles: accountability, responsibility, transparency

AI systems are often characterized by their autonomy, interactivity and adaptability [8, 24]. To reflect societal concerns about the ethics of AI, and ensure that AI systems are developed responsibly, incorporating social and ethical values, we propose to complement these properties with the principles of accountability, responsibility and transparency (ART) [7], as depicted in Fig. 2.

Accountability refers to the need to explain and justify one's decisions and actions to its partners, users and others with whom the system interacts. To ensure accountability, decisions must be derivable from, and explained by, the decision-making algorithms used. This includes the need for representation of the moral values and societal norms holding in the context of operation, which the agent uses for deliberation. Accountability in AI requires both the function of guiding action (by forming beliefs and making decisions), and the function of explanation (by placing decisions in a broader context and by classifying them along moral values).

Responsibility refers to the role of people themselves, and to the capability of AI systems to answer for one's decision and identify errors or unexpected results. As the chain of responsibility grows means are needed to link the AI system's decisions to the fair use of data and to the actions of stakeholders involved in the system's decision.

Transparency refers to the need to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environment, and to the governance of the data used or created. Current AI algorithms are basically black boxes. However, regulators and users demand explanation and clarity about the data used. Methods are needed to inspect algorithms and their results and to manage data, their provenance and their dynamics.

3.1. Responsible AI challenges

In this clause, we discuss how the general principles described above can direct the development of AI systems. Assuming that the development of AI systems follows a standard engineering cycle of Analysis-Design-Implement-Evaluate, taking a design for values approach basically means that the analysis phase will need to include activities for (i) the identification of societal values, (ii) deciding on a moral deliberation approach (e.g. through algorithms, user control or regulation), and (iii) methods to link values to formal system requirements [1].

Responsibility is associated with the capability of moral deliberation, in particular that which is related to dealing with moral dilemmas for which there is not one optimal solution. Several authors have discussed the trolley problem as an example of such a situation. In this scenario, an AI system, e.g. an autonomous vehicle, must decide between harming pedestrians or its own passengers when an accident cannot be avoided. Approaches to moral deliberation reflect ethical theories, such as utilitarianism (save the most lives) or deontological/Kantian (do no harm deliberately). From an implementation perspective, the different ethical theories differ in terms of computational complexity of the required deliberation algorithms. To implement consequentialist agents, reasoning about the consequences of actions is needed, which can be supported by, e.g. dynamic logics. For deontological agents, higher order reasoning is needed to reason about the actions themselves, i.e. the agent must be aware of its own action capabilities and their relations to institutional norms and the rule of law. Accountability requires both the function of guiding action (by forming beliefs and making decisions), and the function of explanation (by placing decisions in a broader context and by classifying them along moral values). To this effect, machine learning techniques can be used to classify states or actions as ‘right’ or ‘wrong’, basically in the same way as classifiers learn to distinguish between cats and dogs. Another approach to develop explanation methods is to apply evolutionary ethics [2] and structured argumentation models [17].

This moreover provides a model-agnostic approach potentially able to deal with transparency in stochastic, logic and data-based models in a uniform way. Further research is needed to verify this approach. Yet another approach is proposed in [12] based on pragmatic social heuristics instead of moral rules or maximization principles. This approach takes a learning perspective integrating both the initial ethical deliberation rules with adaptation to the context. Finally, poorly understood behavior by AI systems can have large and lasting consequences, and adaptive systems may arrive at “perverse instantiations” of their programmed goals [3].

4. CONCLUDING REMARKS

Increasingly, AI systems will be taking decisions that affect our lives and our way of living in smaller or greater ways. In all areas of application, AI must be able to take into account societal values, moral and ethical considerations, weigh up the respective priorities of values held by different stakeholders and in multicultural contexts, explain its reasoning, and guarantee transparency. As the capabilities for autonomous decision making grow, perhaps the most important issue to consider is the need to rethink responsibility. Being fundamentally tools, AI systems are fully under the control and responsibility of their owners or users. However, their potential autonomy and capability to learn, require that design considers accountability, responsibility and transparency principles in an explicit and systematic manner. The development of AI algorithms 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 techniques and applications.

As AI systems replace people in many traditional jobs, it is necessary to 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 fulfillment 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.

Increasingly, robots and intelligent agents will be taking decisions that can affect our lives and way of living in smaller or greater ways. Being fundamentally artifacts, AI systems are fully under the control and responsibility of their owners or users. However, developments in autonomy and learning are rapidly enabling AI systems to decide and act without direct human control. That is, in dynamic environments, their adaptability capabilities can lead to situations in which the consequences of their decisions and actions will not be always possible to direct or predict.

More than being a risk to human values, AI brings in itself enormous potential to improve the lives of many, and to ensure human rights to all. However, how this will be realized, depends on us.

ACKNOWLEDGEMENT

The author is grateful to Thomas van Biemen who performed the study described in clause 2.

REFERENCES

[1] H Aldewereld, V Dignum, and YH Tan. Design for values in software development, 2015.

[2] Ken Binmore. Natural justice. Oxford University Press, 2005.

[3] Nick Bostrom. Superintelligence: Paths, dangers, strategies. OUP Oxford, 2014.

[4] Erik Brynjolfsson and Andres McAfee. The second machine age: Work, progress, and prosperity in a time of brilliant technologies. WW Norton & Company, 2014.

[5] Erik Brynjolfsson, Andrew McAfee, and Michael Spence. Labor, capital, and ideas in the power law economy. Foreign Aff., 93:44, 2014.

[6] Deloitte. From brawn to brains, the impact of technology on jobs in the UK. Technical report, 2015. Accessed: 15-08-2017.

[7] Virginia Dignum. Responsible autonomy. In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJ-CAI'2017), pages 4698-4704, 2017.

[8] Luciano Floridi and Jeff W Sanders. On the morality of artificial agents. Minds and machines, 14(3):349-379, 2004.

[9] Carl Benedikt Frey and Michael A Osborne. The future of employment: how susceptible are jobs to computerisation? Technological Forecasting and Social Change, 114:254-280, 2017.

[10] Batya Friedman, Peter Kahn, and Alan Borning. Value sensitive design and information systems. Advances in Management Information Systems, 6:348-372, 2006.

[11] Dolores Gallego and Salvador Bueno. Exploring the application of the Delphi method as a forecasting tool in information systems and technologies research. Technology Analysis & Strategic Management, 26(9):987–999, 2014.

[12] Gerd Gigerenzer. Moral satisficing: Rethinking moral behavior as bounded rationality. Topics in cognitive science, 2(3):528–554, 2010.

[13] R Gorden. The demise of US economic growth: Restatement, rebuttal and reflections. Technical Report NBER Working Paper No. 19895, National Bureau Economic Research, 2014.

[14] James J. Hughes. A strategic opening for a basic income guarantee in the global crisis being created by AI, robots, desktop manufacturing and biomedicine. Journal of Evolution & Technology, 24(1):45–61, 2014.

[15] K. K. Lilja, K. Laakso, and J. Palomäki. Using the delphi method. In 2011 Proceedings of PICMET '11: Technology Management in the Energy Smart World (PICMET), pages 1–10, July 2011.

[16] McKinsey. Manufacturing the future. Technical report, 2014.

[17] Sanjay Modgil and Henry Prakken. A general account of argumentation with preferences. Artificial Intelligence, 195:361–397, 2013.

[18] Int. Federation of Robotics. The impact of robots on productivity, employment and jobs. Technical report, 2017. Accessed: 15-08-2017.

[19] Chitu Okoli and Suzanne D Pawlowski. The Delphi method as a research tool: an example, design considerations and applications. Information & management, 42(1):15–29, 2004.

[20] Committee on Technology National Science, Technology Council, and Penny Hill Press. Preparing for the Future of Artificial Intelligence. CreateSpace Independent Publishing Platform, 2016.

[21] Thomas Piketty. Capital in the twenty-first century. Harvard University Press, 2017.

[22] Adriano Bernardo Renzi and Sydney Freitas. The Delphi method for future scenarios construction. Procedia Manufacturing, 3:5785–5791, 2015.

[23] Jeremy Rifkin. The Zero Marginal Cost Society. The Internet of Things, the Collaborative Commons, and the Eclipse of Capitalism. Palgrave Macmillan, 2014.

[24] SJ Russell and P Norvig. Artificial Intelligence: A Modern Approach. Prentice Hall, 2nd ed, 2002.

[25] A Smith and J Anderson. AI, robotics, and the future of jobs. Pew Research Center, 6, 2014.

[26] Jeroen van den Hoven. Design for values and values for design. Information Age +, Journal of the Australian Computer Society, 7(2):4–7, 2005.

[27] Quirien van Est and Linda Kool. Werken aan de robotsamenleving: visies en inzichten uit de wetenschap over de relatie technologie en werkgelegenheid. 2015. in Dutch.

[28] John Van Reenen. Employment and technological innovation: evidence from UK manufacturing firms. J Labor Economics, 15(2):255–284, 1997.