AI Communications 31 (2018) 485–494
DOI 10.3233/AIC-180607
IOS Press

The Barcelona declaration for the proper development and usage of artificial intelligence in Europe

Luc Steels [a,*] and Ramon Lopez de Mantaras [b]

[a] *Institución Catalana de Investigación y Estudios Avanzados (ICREA), Institut de Biologia Evolutiva (UPF/CSIC), Barcelona, Spain*
E-mail: steels@arti.vub.ac.be

[b] *Institut d'Investigacio en Intel-ligencia Artificial (CSIC) (IIIA), Barcelona, Spain*
E-mail: mantaras@iiia.csic.es

Abstract. The rapidly increasing deployment of AI raises societal issues about its safety, reliability, robustness, fairness and moral integrity. This paper reports on a declaration intended as a code of conduct for AI researchers and application developers. It came out of a workshop held in Barcelona in 2017 and was discussed further in various follow up meetings, workshops, and AI schools. The present publication is a matter of historical record and a way to publicize the declaration so that more AI researchers and developers can get to know it and that policy makers and industry leaders can use it as input for governance. It also discusses the rationale behind the declaration in order to stimulate further debates.

Keywords: Ethical issues of AI, Barcelona declaration

1. Motivation

1.1. The AI summer is here

It can no longer be denied that Artificial Intelligence is having a fast growing impact in many areas of human activity. It is helping humans to communicate with each other – even beyond linguistic boundaries, find relevant information in the vast information resources available on the web, solve challenging problems that go beyond the competence of a single expert, enable the deployment of autonomous systems, such as self-driving cars or other devices that handle complex interactions with the real world with little or no human intervention, and many other useful things. These applications are perhaps not like the fully autonomous conscious intelligent robots that science fiction stories have been predicting, but they are nevertheless very important and useful, and most importantly they are real and here today.

The growing impact of AI has triggered a kind of 'gold rush': we see new research laboratories springing up, new AI start-up companies, and very significant investments, particularly by big digital tech companies, but also by transportation, manufacturing, financial, and many other industries. Management consulting companies are competing in their predictions how big the economical impact of AI is going to be and governments are responding with strategic planning to see how their countries can avoid staying behind.

Clearly most of the activity is in the US [18] and China but there are also signs of enhanced AI activity in Europe and anouncements of action plans by various European governments and the European Commission. The Macron 1.5 billion Euro strategic plan for stimulating AI in France [26] is one example. Although European strategic proposals are today (i.e. in 2018) mostly still in the phase of promises, European AI researchers, developers and entrepreneurs hope that they will provide structural funding for AI in the near future and that AI becomes recognized in upcoming European framework programs as a research field with a clear economic impact and hence in need of significant structural funding.

*Corresponding author. E-mail: steels@arti.vub.ac.be.

1.2. Clouds on the horizon

Although all this is positive news, it cannot be denied that the application of AI comes with certain risks. Many people (including luminaries such as Bill Gates, Elon Musk, or Stephen Hawking) believe that the main risk of AI is that its deployment would get out of hand. Machines that can learn, reconfigure themselves, and make copies of themselves may one day outrun the human race, become smarter than us and take over. To researchers in the field this risk seems far-fetched. But they see other risks, which are already upon us and need urgent remediation. Here are some examples:

Example 1. AI algorithms, particularly those embedded in the web and social media, are having an important impact on who talks to whom, how information is selected and presented, and how facts (justified or fake) propagate and compete in public space. Critics point out that these AI algorithms are now held (at least partly) responsible for allowing the emergence of a post-truth world, highjacking democratic decision processes, and dangerously polarizing society. Polarization is making it much more difficult to deal with the big issues facing our society, such as climate change mitigation, diminishing pollution, achieving economic prosperity for an exploding world population, avoiding violent conflicts due to ethnic, nationalistic or religion diversity, coping with massive migration, etc. They all require determined collective action and therefore a political consensus. AI should (and could) help to support consensus formation rather than destroy it.

Example 2. Many applications use deep learning or other forms of statistical inference to great advantage. For many of these applications, such as speech recognition or machine vision, this technique is the most effective one found so far. But the applications of deep learning to domains that involve rule-governed behavior and human issues, such as financial decision making, human resource management, or law enforcement, has been shown to be problematic from a humanistic point of view. Job seekers report frustration to get through the machine learning based filters which reinforce gender and class and focus on keywords or features of a cv that are not essential nor fair [12]. The use of AI in decisions on parole has caused an outcry because the basis of these decisions is obscure due to the black box nature of deep learning and biased in ways that are unacceptable [6]. All of this raises growing questions about the robustness, explainability, reliability and accountability of AI systems based on deep learning.

Example 3. Self-driving cars act upon their own decisions, unavoidably leading to risks to human life. More generally, do we need to put limits on autonomous artificial intelligence systems? Who is responsible when something goes wrong? And what about other applications of autonomous AI such as autonomous weapons. The AI community is already speaking out against their use[1] but without a world-wide consensus against their deployment, as in the case of chemical weapons or landmines, and reluctance or refusal of AI developers to participate in their development, we risk a new kind of arms race or a risk that violence is used more easily to resolve conflicts.

1.3. The Barcelona initiative

Several initiatives have been taken in recent years to understand better the risks of AI deployment and come up with legal frameworks, codes of conduct, and value-based design methodologies. Examples are the Alomar principles for beneficial AI,[2] the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems,[3] the technology industry consortium 'Partnership on AI to benefit people and society',[4] or the EU GDPR regulation which includes the right for an explanation [10]. There is also a rapidly growing literature on the risks of AI and how to handle it (see for example [3–5,17,21]). hybrid Against this background, Luc Steels and Ramon Lopez de Mantaras organised in march 2017 a debate in CosmoCaixa Barcelona under the auspices of the Biocat and l'Obra Social la Caixa with support from ICREA, the Institut de Biologia Evolutiva (UPF/CSIC) and the Institut d'Investigacio en Intel-ligencia Artificial (CSIC). More information about this event is found here: http://www.bdebate.org/en/forum/artificial-intelligence-next-step-evolution

The event assembled a number of top experts in Europe who are concerned with the benefits and risks of AI – particularly, but not exclusively, in the domain of web and social media – and to discuss strategies to deal with these risks. The Barcelona initiative is complementary to other ongoing efforts because (i) it intends to stimulate the debate within Europe whereas other initiatives are primarily in the Anglo–American sphere, and (ii) give a voice to European AI developers

[1] See for example https://futureoflife.org/open-letter-autonomous-weapons/.

[2] https://futureoflife.org/ai-principles/

[3] https://standards.ieee.org/develop/indconn/ec/auto-sys-form.html

[4] https://www.partnershiponai.org/

and researchers, whereas most of the discussion on ethical AI so far has been dominated by social scientists, legal experts and business consultancy firms.

The Barcelona meeting featured a small-scale workshop on March 7th with two sessions followed by discussion:

- **Session 1** raised the question: Is AI ready for large-scale deployment? AI algorithms are now being used on a grand scale for applications ranging from news selection, medical diagnosis, insurance, self-driving cars, etc. But is the current technical state of the art ready for these challenges? What new research needs to be done to make AI-based systems more accountable and trustworthy? The discussion was kickstarted with contributions from Marcello Pelillo (Director European Centre for Living Technologies, University of Venice) and Hector Geffner (Head of the Artificial Intelligence Group, DTIC, Universitat Pompeu Fabra Barcelona).
- **Session 2** focused on the societal impact of AI. AI is now used primarily for commercial purposes, but can we also use AI for the common good? What applications should we encourage? How can such development be financed? Many commercial applications now have a strong manipulative character and ignore privacy considerations in order to get sufficient data for statistical learning. We also see increasing usage in political propaganda which potentially endangers healthy democratic decision-making. How can these negative effects in the deployment of AI be addressed? This discussion was kickstarted with contributions from Camilo Cristancho (Professor Political Science, Universitat Autonoma Barcelona) and Antoni Roig (Professor of constitutional law, Institute of Law and Technology, Universitat Autonoma de Barcelona).

The second day was open to the public. It featured the following presentations, recorded and available through: http://www.bdebate.org/en/videos

Session 1. Dreams – How is AI presented in popular culture? AI through the eyes of cinema and literature. Carme Torras (IRI – CSIC/UPC).

Session 2. Reality – What are recent technical breakthroughs in AI and how do they impact applications?

- Part A. Advances in knowledge-based AI:
  1. How is the semantic web transforming information access. Guus Schreiber (Network Institute, Vrije Universiteit Amsterdam)
  2. How are advances in language processing helping to bring order in cyberspace. Walter Daelemans (Computational Linguistics, University of Antwerp).
- Part B. Advances in machine learning:
  1. How do recent developments in deep learning increase its power and scope of application. Joan Serra (Telefonica I+D, Barcelona),
  2. Why is the industrial impact of machine learning growing so fast? Francisco Martin (BigML, Corvallis Oregon US).

Session 3. The role of AI in social media.

- How do AI algorithms influence the selection of media content. Cornelius Puschmann (Hans Bredow Institute for Media Research, Hamburg)
- The complex dynamics of rumour and fake news spreading. Walter Quattrociocchi (Ca'Foscari University of Venice)

Session 4. Making AI safe and beneficial

- Best practices for the development and deployment of AI. Francesca Rossi (University of Padova, Italy)
- Technologies for the democratic city. Francesca Bria. (Comisionada de TecnologÌa e Innovacion Digital, Ayuntamiento Barcelona)

The symposium concluded with a panel discussion and a presentation by Luc Steels of the 'Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe'. The declaration was then signed by most of the participants and from then on has become accessible for signature and discussion on the web.

2. The declaration

This section reprints the complete text of the declaration. A summary is provided in Table 1.

Barcelona declaration for the proper development and usage of artificial intelligence in Europe

1. Scope. AI is a collection of computational components to build systems that emulate functions carried out by the human brain. The field started in the mid-nineteen fifties and has since gone through cycles of promise, enthusiasm, criticism and doubt. At this moment we see a strong wave of enthusiastic adoption of AI in many areas of human activity.

We distinguish between knowledge-based AI and data-driven AI.

Table 1
Main points of the Barcelona Declaration

1.	Scope: We consider both knowledge-based and data-driven AI
2.	Investment: Europe needs to scale up its AI effort.
3.	Prudence: Honest communication is needed about the strenghts and limitations of AI applications.
4.	Reliability: It is necessary to create a trusted organisation for verification and validation of AI.
5.	Accountability: AI applications need to be intelligible, able to explain the basis of their decisions.
6.	Identity: It should always be clear whether we are dealing with a human or an AI system.
7.	Autonomy: Rules need to be found to constrain autonomous behavior.
8.	Maintaining human knowledge: Human intelligence needs to be fostered as a source of future knowledge.

- **Knowledge-based AI**, which has become applicable in the late seventies, attempts to model human knowledge in computational terms. It starts in a top-down fashion from human self-reporting of what concepts and knowledge rules individuals use to solve problems or answer queries in a domain of expertise, including common sense knowledge, and then formalizes and operationalizes this as software components. Knowledge-based AI emphasizes conceptual models, ontologies, common sense knowledge bases, reasoning and problem solving strategies, language processing, and insight learning. It rests primarily on highly sophisticated but now quite standard symbolic computing technologies.
- **Data-driven AI**, also commonly known as machine learning, became applicable only the last decade. It starts in a bottom-up fashion from large amounts of data of human activity, which are processed with statistical machine learning algorithms, such as the deep learning algorithm, in order to abstract patterns that can then be used to make predictions, complete partial data, or emulate behavior based on human behavior in similar conditions in the past. Data-driven AI requires big data and very substantial computing power to reach adequate performance levels.

Knowledge-based AI has shown to be most successful in intellectual tasks, such as expert problem solving, whereas data-driven AI is most successful in tasks requiring intuition, perception, and robotic action. The full potential of AI will only be realized with a combination of these two approaches, meaning a form of **hybrid AI** [12].

2. Investment. The current surge of interest and application of artificial intelligence (AI) is without precedent. There is a growing consensus that AI is of huge importance for the future economy and functioning of European society. AI is now understood to be a powerful, novel way to link producers and consumers, and a novel way to add value to products, build new ones, and improve production processes. Moreover AI can help to introduce more efficiency and quality into bureaucratic procedures and give greater access to knowledge and creativity for all. We therefore call upon European funding agencies and companies to invest in the development of AI at a scale which is adequate for the challenge, and in such a way that ALL European regions and citizens can profit. This investment should target the creation of a complete ecosystem with a network of high-end research labs with sufficient structural (as opposed to project-based) funding, diffusion of AI techniques to form a significant number of 'AI engineers', and proper conditions and stimuli for successful AI entrepreneurship. Of particular importance is the development of open resources, such as corpora, ontologies and software frameworks, that should be available as the common infrastructure on which specific applications get built. Because many of these resources are specific to individual languages and cultures, it is important that Europe invests in their development, partly to make applications accessible and adapted to all European regions. Europe currently lags behind other economic areas in the investment in AI and the time for a very significant scale-up is now.

3. Prudence. The leap forward in AI has been caused by a maturation of AI technologies, vastly increased computing power and data storage, the availability of delivery platforms through the internet, and an increased willingness of many economic actors to try out the technology for their own application domain. But we must be aware of the limitations. Many fundamental problems of artificial intelligence are not yet solved and will require radical breakthroughs. Solving them will require substantial long-term research efforts. The application of AI also demands very stringent prerequisites; otherwise the results will be disappointing and potentially very harmful. For example, the application of knowledge-based AI requires

the availability of human expertise and sufficient resources to analyze and model it. The application of data-driven AI requires enough high quality data and careful choices of which algorithms and parameters are appropriate in each case. These application prerequisites need to be investigated and spelled out in much more detail so that those responsible for applying AI can exercise the necessary prudence.

4. Reliability. All artificial systems that are used in our society have to undergo tests to determine their reliability and security. So it is normal that the same is done for AI systems, particularly in domains like medicine or autonomous robots. Although verification and validation procedures have been developed for knowledge-based systems in the nineteen-eighties and nineties, they are still lacking for data-driven AI. Sure, at the moment machine learning practices make a distinction between an example data set used for training and a test set used to gauge in how far a system has reached adequate levels of performance, but there is still a significant difference between a test set and actual testing in real world conditions. Moreover, once adequate verification and validation methodologies are available, we will need a network of agencies in European countries (or a central European agency) that use them. They should become the authority to validate AI applications before they are put into widespread usage. The European Parliament has recently decided to create an agency for robotics and AI and this agency could potentially take up this task.

5. Accountability. When an AI system makes a decision, humans affected by these decisions should be able to get an explanation why the decision is made in terms of language they can understand and they should be able to challenge the decision with reasoned arguments. This is particularly important in domains such as decisions on loans, legal decisions (for example about granting parole[5]), insurance, taxation, etc. AI systems, particularly those based on data-driven approaches, are currently unable to provide this kind of explanation. Their decisions are based on a large set of statistically derived network parameters. Techniques to make sense of these parameters are in their infancy and will probably require a combination of knowledge-based and data-driven AI. Nevertheless we should not allow widespread application usage without a solution to the accountability problem, and accountability should become a precondition for deployment.

[5] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

6. Identity. There is a growing worry about AI chatbots or other kinds of automatic messaging systems operating on the Internet and in social media, designed for the manipulation of political opinion, disinformation through the propagation of false facts, extortion, or other forms of malicious activity that is dangerous for individuals and destabilizing our society. These chatbots pretend to be human and do not give away the identity of those behind them. The use of AI has made these chat-bots sufficiently realistic that unsuspecting users are not able to make a distinction and get mislead. A possible solution to this issue is to demand that it is always clear whether an interaction originates from a human or from an AI system, and that, in the case of an artificial system, those responsible for it can be traced and identified. This solution could possibly be implemented by a system of water marking and become mandatory in Europe.

7. Autonomy. AI systems have not only the capacity to make decisions. When they are embedded in physical systems, such as self-driving cars, they have the potential to act upon their decisions in the real world. This understandably raises questions about safety and about whether autonomous AI will not at some point overtake humans. Although some of these worries belong more in the domain of science fiction than reality, the proper circumscription of autonomous intelligent systems is an important challenge that must be addressed. It is necessary to have clear rules constraining the behavior of autonomous AI systems, so that developers can embed them in their applications. It is also necessary to clarify who is responsible for failure – as is indeed the case with all products.

8. Maintaining Human Knowledge. The undeniable enthusiasm for AI gives sometimes the impression that human intelligence is no longer needed and it has lead some companies to fire employees and replace them by AI systems. This is a very serious mistake. All AI systems critically depend on human intelligence. Knowledge-based systems model the knowledge and insight of human expertise and data-driven AI systems rely critically on data of human behavior. It follows that human expertise should continue to be taught, developed and exercised. Moreover in almost any area, human expertise still far outstrips artificial intelligence, particularly for dealing with cases that have not appeared in the example data sets from which AI systems are learning.

We believe that AI can be a force for the good of society, but that there is a sufficient danger for inappropriate, premature or malicious use to warrant the need

for raising awareness of the limitations of AI and for collective action to ensure that AI is indeed used for the common good in safe, reliable, and accountable ways.

Barcelona, 8 March 2017

The list of current signataries is available through the website:

https://www.iiia.csic.es/barcelonadeclaration/

It is still possible to sign the declaration through the same site.

3. Follow up

After the event in Barcelona, the declaration was spread through various AI research channels and public media. It was integrated in various discussions on the future governance of AI in Europe, for example, at a hearing in Brussels of the EU political Strategy Center.[6] The declaration was also discussed at various AI schools and fora, such as the 2017 edition of the Interdisciplinary College IK in Guenne, Germany.

In general, the declaration contributed to raise awareness and has given additional impetus to initiatives by governments and law makers in many European countries, such as the Netherlands [20], Belgium [23], Denmark [24], the UK [1], a.o. In some cases, the recommendations of the declaration were explicitly referred to in parliamentary hearings [13]. In addition, the European Commission initiated in the spring of 2018 a High-level Expert Group on Artificial intelligence[7] as a steering group of the newly formed European AI alliance,[8] which includes stakeholders ranging from industry to policy makers and academics.

So, although the landscape of AI in Europe is rapidly changing through all these discussions and activitites, the issues raised in the declaration remain highly relevant. The remainder of this section highlights some of them.

1. There is an even greater need today to clarify what we mean by AI when discussing legal and ethical issues. The first item of the declaration was intended to circumscribe the field. This has become even more urgent. At the moment AI is increasingly being used as an umbrella term for a wide range of techniques that used to be classified under operations research, pattern recognition, information retrieval, data analytics, business modeling, statistical analysis, etc. The term data science is currently used for these activities and is indeed much more appropriate. AI is also used for developments in digital media, particularly social media, even though its role is in many cases nonexistent.

When AI is interpreted too broadly like this, there is a risk, for the field of AI itself, to be blamed for malicious applications, business practices or societal phenomena, such as fake news, hate speech or cyber crime, even if no AI is involved at all. The declaration therefore proposed to focus only on ethical issues as related to AI in the narrow sense. And even if we maintain this restriction, we need to be more precise whether we are talking about knowledge-based AI or data-oriented machine learning, because the legal and ethical issues for both approaches are quite different. For example, the topic of explainability was already an important component of knowledge-based systems built in the 1980's and adequate approaches have been developed and used extensively [16], whereas explanation is highly problematic for current machine learning techniques such as deep learning and it is still very unclear how it could be achieved [2].

2. The plans for supporting the development and deployment of AI still have to become concrete in Europe. Since march 2017, the call for greater and above all more stable investment in European AI, as proposed in the second recommendation of the Barcelona declaration, has been 'heard' in several European countries, giving rise to considerable optimism. The Macron 1.5 billion Euro action plan for AI in France [26] has already been mentioned, but there are also plans being drawn by the Merkel government in Germany (anounced for autumn 2018), and by several smaller countries. Moreover in april 2018 most of the member states of the European Union have signed a Declaration of Cooperation on Artificial Intelligence (AI) in order to combine and coordinate their efforts.[9]

All of these initiatives are very welcome but they are statements of intent and concrete actions with a direct impact on the deployment of AI, or, just as important, on the creation of stable funding for AI research and

[6] https://ec.europa.eu/epsc/events/high-level-hearing-european-union-strategy-artificial-intelligenceen

[7] https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

[8] https://ec.europa.eu/digital-single-market/en/european-ai-alliance

[9] https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence

education in Europe are still rare. One positive sign was a first 20 mi EU call within the European H2020 framework program (with deadline march 2018) that explicitly targeted the stimulation of European AI research in line with what was proposed in the Barcelona declaration, namely the creation of a European ecosystem and a platform in which European actors could share resources in the form of machine learning algorithms, data sets, knowledge bases, ontologies, lexicons, grammar models, etc. The AI4EU consortium has been selected and activity is planned to start in January 2019. Moreover Cecile Huet from the Future and Emergent Technologies office (DG CNECT) outlined the Artificial Intelligence Strategy for Europe at IJCAI-2018 in Stockholm, which foresees an important increase of opportunities for AI research (70%), mainly because a large number of calls will include the possibility to integrate AI. Of particular significance is that now "AI-oriented proposals in any basic or applied research domain are welcome" for the ERC calls. Beyond the Horizon 2020 program, it is foreseen that the next Multi-Annual Financial Framework earmarks 2.5 billion EU for AI.

3. We need more rather than less prudence with respect to what AI can do. The current hype in AI is largely fueled by anouncements and demonstrations by American tech companies, by web-based blogs and news media that amplify expectations or prototypes, by very rapid publishing (for example through blogs or arXiv) without the reviewing process characteristic of normal scientific communication, and, to some extent, by a lack of cognitive science background by newcomers to the field, who appear unaware of the enormous complexity of human knowledge and knowledge processing and therefore heavily underestimate the challenges of 'real' AI. Moreover there is the old AI disease of reading too much into the performance of an AI system, for example assuming that an artificial agent has intentions like deception, whereas there is only the appearance of deceptive behavior to the human observer without any explicit intentional goal to deceive by the agent.

A typical example is a recent hype episode about an experiment carried out by Facebook researchers on the acquisition of skills in negotiation, with the acquisition of language skills as a secondary needed competence [15]. The paper was published on arXiv which ensures rapid dissemination in the machine learning community, and on a company blog,[10] which ensures that the experiment is picked up by the media. So far so good.

However, a report in the media, on the website Fast Company, did not discuss the negotiation experiment itself but focused entirely on the acquisition of language skill with the headline: "AI Is Inventing Languages Humans Can't Understand. Should We Stop It?", commenting "Researchers at Facebook realized their bots were chattering in a new language. Then they stopped it."[11] Indeed, non-English dialogs started to be produced such as this one:

Bob: you i everything else....
Alice: balls have a ball to me to me to me to me to me to me to me.

Although the researchers never mentioned anything about stopping the experiment for this reason.

The phenomenon of novel language emergence is in itself interesting, particularly to those in the AI research community that have been studying for decades the cultural evolution of language through a wide range of agent-based experiments, including with embodied robots [22]. But then, web media, blogs, and newspapers picked up the theme of self-generated language and elaborated only on the potential dangers with stories that became progressively more and more scary. For example, the UK tabloid The Sun reported: "Facebook shuts off AI experiment after two robots begin speaking in their OWN language only they can understand" and quotes Kevin Warwick as "anyone who thinks this is not dangerous has got their hand in the sand". The Sun adds: "The incident closely resembles the plot of The Terminator in which a robot becomes self-aware and starts waging a war on humans."[12] Newspapers all over Europe picked up the story, all adding their own twists and exagerations and confronting AI researchers with this supposed step towards AI disaster and urging politicians to stop this madness.

The Facebook researchers involved surely did not intend this media storm but these stories are the ones that stick into the public understanding of AI. Clearly much greater care needs to be taken in communicating AI experiments. Otherwise the lack of prudence will without doubt lead to a new AI winter as the expectations and scare stories currently being created by overzealous media are impossible to fulfill and they overpower the more reasonable statements that most AI researchers tend to make.

[10] https://code.fb.com/ml-applications/deal-or-no-deal-training-ai-bots-to-negotiate/

[11] https://www.fastcompany.com/90132632/ai-is-inventing-its-own-perfect-languages-should-we-let-it

[12] https://www.thesun.co.uk/tech/4141624/facebook-robots-speak-in-their-own-language/

4. No mechanisms for certification of AI systems have been put in place yet. The European parliament already voted in early february civil law rules on robotics, which also are highly relevant for AI and recommended the European Commission to set up an agency that would be tasked with certification.[13] But certification of AI is easier said than done. There are already some small-scale European initiatives exploring whether specific AI algorithms[14] are performing as they are claimed to be and there are various initiatives for value-based design, some reflecting the conclusions [7].

All this work builds up possible experience that can lead to certification procedures but there is still a considerable road to travel before this will be as uncontroversial as certifying a new refrigerator or new medicine.

5. Accountability through explainable AI and legal personhood. Accountability of AI systems continues to be a cause of great concern. It covers two aspects: being able to get an explanation to understand how a decision was reached, and being able to identify who is ultimately responsible for a mishap. Accountability should not be relegated to regulations to be applied once the AI system has been deployed, instead it must betaken into account at design time. Explainability is necessary not only because transparent explainable AI systems allow users to trust the systems, but also because it is a crucial element for accountability. Indeed, explanations should not just be traces but justifications. A type of justification could be, for instance, a contrastive explanation. That is, answering the question 'why output A instead of output B?'. Relevant explanations allow to inspect why and how an AI system came to its conclusions and to locate possible errors and biases in the design of the systems. Besides being contrastive, explanations should also be selective, that is they should focus on the most relevant features that led to the output. Such rich explainability scenarios will require the capability of counterfactual reasoning [19], and are not only very much needed in data-driven AI, due to its black box nature, but also in knowledge-based AI.

6. The Identity of demonstrated AI systems keeps getting blurred. The sixth recommendation of the declaration argued that AI systems should make it very clear that they are artificial rather than human. The Turing test is in that respect misleading – because it suggests that fooling humans into believing that an artificial system is indistinguishable from a human is the ultimate goal of AI. It is not the goal of AI and should never be. AI can be useful without this kind of deception, particularly because sooner or later human users detect the deceipt anyway.

Here is a recent example where this recommendation has been violated. In may 2018 Google demonstrated a speech understanding system called Google Duplex that is claimed to be able to hold a conversation over the phone for ordering a reservation in a restaurant or make similar appointments for services [14]. This result is quite interesting although the boundaries of the system's performance are not very clear and some have even questioned whether the demo was in real circumstances. But the fact that Google Duplex tried to con humans into believing that the conversation was not with a machine, created an immediate backlash and a promise that in the future the system would identify itself.[15] Moreover some observers were quick to realise the abuse that could be made of this technology. Here is for example a typical reaction (posted on the Google Blog anouncing the Duplex demonstration).

> A couple of problems spring immediately to mind. First, the use of embedded "uh"s and other artifacts to try fool the listener into believing that they are speaking to a human may well engender blowback as these systems are deployed. My sense is that humans in general don't mind talking to machines so long as they know that they're doing so. I anticipate significant negative reactions by many persons who ultimately discover that they've been essentially conned into thinking they're talking to a human, when they actually were not. It's basic human nature – an area where Google seems to have a continuing blind spot. Another problem of course is whether this technology will ultimately be leveraged by robocallers (criminal or not) to make all of our lives even more miserable while enriching their own coffers.

A possible response to avoid these problems is to legally require that any AI system should make it explicitly clear upfront that it is an artificial system, so that human users can also shield themselves from calls by such systems.

7. Most issues related to the autonomy of AI remain open. The question how much autonomy should

[13] http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2017-0051+0+DOC+PDF+V0//EN

[14] https://algorithmwatch.org/en/the-adm-manifesto/

[15] https://www.techspot.com/news/74582-google-responds-duplex-backlash-ai-voice-system-identify.html

be given to an AI system is for many applications, such as weapon technology or autonomous cars, of primordial importance. There are two avenues with which this issue is being approached. The first one is to create rules of governance and a legal framework that is both a guideline for developers and a mechanism by which those impacted negatively by the technology can seek redress. There is a considerable amount of recent work in this area with very specific proposals being developed in several European countries. For example, the German Ministry for Transport and Digital Infrastructure has already issued a set of ethical guidelines for the design and deployment of autonomous cars which is now legally binding [8]. The main points are: (i) Autonomous driving is ethically not only justified but obligatory because it can lead to a general decrease in the number of accidents. (ii) Saving human life has always a higher priority than avoiding material damage. (iii) There should be no discrimination with respect to saving life. (iv) It must always be clear who is responsible, human or machine. (v) All data must be stored to clear up future misunderstandings and these data must be accessible and under the control of the driver. A similar initiative in France is expected to issue guidelines by 2019. But it would obviously be better if there is a European initiative on this matter to avoid divergences between member states. The second avenue is to integrate moral decision-making and legal rules in the behavior of the AI system itself. Here also there is a considerable history of prior discussion [27] and occasional technical work. [9] that requires however much deeper exploration before being usable in practice.

8. Maintaining Human Knowledge. AI systems critically depend on human intelligence but also humans can greatly benefit when teaming up with AI systems. The development of AI should be human-centered, that is we should shift the focus from machines replacing human workers to tools, assistants, and in the long term peers that will help, complement and leverage humans in performing tasks and taking decisions leading to better results and higher quality of jobs. To achieve the 'peer' level of collaboration, humans and machines will have to share goals and cooperate synergistically towards their fulfillment. Synergy is due to the complementary strengths of humans and machines. Humans, for instance, are much better than machines at adapting to unforseen changes when performing a task and dealing with unexpected and uncertain situations in general. AI systems are better than humans in aspects such as recognizing patterns, memorising and analysing large amounts of data and information. There are already many examples that show the synergetic advantage of human-machine coworking: An excellent example is the work Combining Deep Learning with a human pathologist for Identifying Metastatic Breast Cancer [28]. In this study, based on several hundred cases, a top human pathologist achieved a percentage of error of 3.4%, a deep learning system an error of 7.5% but the combination of both reduced the error to only 0.52%. Therefore, the emphasis should be on how machines and humans can be co-workers instead of machines replacing humans.

On the other hand, the discussion on automation and employment is erroneously centered on only the number of jobs. Instead it should be focused on the changing nature of work. Some tasks have been, and will continue to be, automatized but the number of jobs where the majority, or all, of its tasks can be automatized is not as large as some studies say. There are studies in different European countries that show how robotization has in fact increased the global number of jobs, creating jobs with higher quality and better paid. On example is a study in Catalonia done at the Catalan Open University with several hundred SMEs from 2002 to 2014 [25].

A similar study done by the Centre for European Economic Research in Mannheim, Germany, also shows that automation resulted in an overall increase in (better paid) jobs in Germany between 2011 and 2016. This does not mean that we should not be concerned by the effects of AI and automation in general on employment (not only robots in manufacturing but also clerical work and professional services) but we should focus our concerns and find solutions to the problem of the changing nature of jobs, train workers to face this challenge and ensure that automation does not increase inequality in society.

4. Conclusions

This paper contributes to ongoing discussions in Europe related to the ethical issues of AI. We focused on the 'Barcelona Declaration for the Proper Development and Use of AI', which was launched in the spring of 2017, and discussed some of its ramifications. Given the public interest in AI and the eagerness of many organisations, both private companies and governmental institutions, to develop applications that affect people in their daily lives, it is important that the AI community, encompassing application developers as well

as researchers, engages in open discussions, partly to avoid over-expectations with an unavoidable backlash later and partly to avoid improper usage of AI that causes unneeded negative side effects and undue human suffering. At the same time, we must realize that no set of rules or in-built technological constraints can ever avoid malicious use by unscrupulous actors. The ultimate responsibility always lies with humans, both as designers and as users, and they should be held accountable.

References

[1] J. Bakewell et al., AI in the UK, ready, willing and able? House of Lords, Select Committee on Artificial Intelligence, Report 2017-19, https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

[2] O. Biran and C. Cotton, Explanation and justification in machine learning: A survey proc, in: *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*, 2017, pp. 8–13.

[3] P. Boddington, *Towards a Code of Ethics for Artificial Intelligence*, Springer-Verlag, 2017.

[4] N. Bostrom, *Superintelligence. Paths, Dangers, Strategies*, Oxford University Press, Oxford, 2016.

[5] M. Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Report Future of Life*, Institute, Oxford, 2018.

[6] A.J.B Caliskan and A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, *Science* **356**(6334) (2017), 183–186. doi:10.1126/science.aal4230.

[7] V. Dignum, Responsible artificial intelligence: Designing AI for human values, *ICT Discoveries* **1**(1) (2018), 1–8.

[8] Ethik-Kommission BMVI, Automatisiertes und Vernetztes Fahren. Eingesetzt durch den Bundesminister für Verkehr und digitale Infrastruktur, 2017, https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf?==blob=publicationFile.

[9] A. Etzioni and O. Etzioni, Incorporating ethics into AI, *The Journal of Ethics* **21**(4) (2017), 403–418. doi:10.1007/s10892-017-9252-2.

[10] B. Goodman and S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", 2016, https://arxiv.org/abs/1606.08813.

[11] Goyal et al., Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, CFPR, 2017.

[12] S. Kanthal, Algorithm overlords, BBC podcast, 2018, https://www.bbc.co.uk/programmes/b0b4zxcn.

[13] C. Lacroix et al., Session 24 May 2018 of Belgian Senate to start investigation on impact, opportunities and risks of the digital smart society, Belgian Senate Official Publication, 2018.

[14] Y. Leviathan and Y. Matias, Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone. Google AI Blog, 2018, https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html.

[15] M. Lewis, D. Yarats, Y. Dauphin, D. Parikh and D. Batra, Deal or No Deal? End-to-End Learning for Negotiation Dialogues, 2017, https://arxiv.org/abs/1706.05125v1.

[16] J. Moore and W. Swartout, Explanation in Expert Systems: A survey, USC/ISI RR-88-288, 1988.

[17] V. Mueller (ed.), *Risks of Artificial Intelligence*, CRC Press, Chapman and Hall, London. doi:10.1201/b19187.

[18] L. Parker, Creation of the national artificial research and development strategic plan, *AI Magazine* **39**(2) 25–32. doi:10.1609/aimag.v39i2.2803.

[19] J. Pearl and D. Mackenzie, *The Book of Why*, Basic Books, New York, 2018.

[20] Rathenau Institute, Acties voor een verantwoorde digitale samenleving. Bericht aan het Nederlandse Parlement, 2018, https://www.rathenau.nl/nl/publicatie/bericht-aan-het-parlement-acties-voor-een-verantwoorde-digitale-samenleving.

[21] S. Russell, Ethics of artificial intelligence, *Nature* **521**(7553) (2015), 415–416. doi:10.1038/521415a.

[22] L. Steels (ed.), *Experiments in Cultural Language Evolution*, John Benjamins Pub., Amsterdam, 2012.

[23] L. Steels (ed.), *Artificiele Intelligentie. Naar Een Vierde Industriele Revolutie?* Royal Flemish Academy of Belgium for Arts and Sciences, Brussels, 2018, http://www.kvab.be/sites/default/rest/blobs/1489/nw-artificieleintelligentie.pdf.

[24] M. Steensen et al., *Artificial Intelligence in Denmark*, Microsoft, Denmark, 2018, https://news.microsoft.com/uploads/prod/sites/53/2018/04/Report-Artificial-Intelligence-in-Denmark-potentials-and-barriers.pdf.

[25] J. Torrent-Sellens and A. Diaz-Chao, Coneixement, robòtica i productivitat a la PIME industrial catalana: Evidència empírica multidimensional. Congrès Català d'Economia, 2018.

[26] C. Villani et al., Donner un sens a l'intelligence artificielle, Pour une strategie Nationale et Europeeene, Mission Parlementaire, France, http://aiforhumanity.fr.

[27] W. Wallach and C. Allen, *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, Oxford, 2009.

[28] D. Wang, A. Khosla, R. Gargeya, H. Irshad, A.H. Beck and B. Israel, Combining Deep Learning with a human pathologist for Identifying Metastatic Breast Cancer, 2016, arXiv:1606.05718v1.