ALLAI.

ALLAI | AIREA-NL

Recommendations for the Artificial Intelligence Research Agenda for the Netherlands

Mr. Catelijne Muller

September 25, 2019

ALLAI | AIREA-NL

ALLAI recommendations for the Artificial Intelligence Research Agenda for the Netherlands

Mr. Catelijne Muller[1]

NWO (the Netherlands Organisation for Science Research) is currently drafting the Artificial Intelligence Research Agenda for the Netherlands (AIREA-NL). NWO has requested ALLAI to provide input and give recommendations for the first AIREA-NL "Version for field consultation" of September 2, 2019.

Below are ALLAI's recommendations regarding a number of Chapters/Subchapters of the AIREA-NL Version for field consultation (1. Scope and Urgency, 2. Motivation and Structure, 3. Grand AI Research Challenges and 5. Impact of AI in Application Domains). ALLAI also suggests to add a number of additional research questions (RQ) to AIREA-NL. Each recommendation or additional RQ is followed by a motivation.[2]

1. SCOPE AND URGENCY

- ***ALLAI recommends to stay away from the AI-race discourse***

In the chapter on 'Scope and Urgency' of the "Version for field Consultation" it is assumed that there is an AI-race ongoing (using wording such as 'winner-takes-all, 'geopolitical race for power'), which would bring urgency to a more favourable AI research climate. In ALLAI's opinion, the 'AI-race discourse' is both wrong and dangerous.[3] It is far more important that AI is being developed and used responsibly, grounded in ethical principles and human rights. This is not a burden on research nor does it stifle innovation, but is the stepping stone that will bring this powerful technology forward. More than a technical decision, this is one of policy and vision, in which Europe is currently leading the way. Besides, it is often forgotten that Europe is the largest market. This gives the EU the opportunity to set the preconditions for AI and thereby also have global influence.

2. MOTIVATION AND STRUCTURE

- ***ALLAI recommends to add to the need for collaboration between research domains to this chapter***

AI can no longer be seen as merely a technical or computer science discipline. It is by definition multidisciplinary. This is already being acknowledged in the draft research agenda. In addition to that, collaboration between research domains is crucial.[4]

[1] President ALLAI, Member of the High Level Expert Group on AI to the Europese Commissie (AI HLEG), President Thematic Study Group on AI of the Economisch en Sociaal Comité (EESC), Rapporteur AI for the EESC.

[2] Due to the method of delivery, the original recommendations sent to NWO did not contain footnotes.

[3] Virginia Dignum, "There is no AI – race and if there is, it's the wrong one to run".

[4] Opinion on AI & Society by the European Economic and Social Committee (Rapporteur Catelijne Muller), INT/806.

Apart from separately researching the legal, ethical and societal impact of current AI-systems, a collaborative approach should be chosen for further development of AI, whereby legal, ethical and social effects are being researched alongside the research into a particular AI innovation.

For funding purposes this means that in parallel with funding for the development of disruptive of AI innovations, there should also be a substantial part of such funding allocated to research into the societal impact of these particular innovations innovations and ways of addressing them.

A good example of an existing multidisciplinary collaborative approach is the The Research Priority Area (RPA) Human(e) AI at the University of Amsterdam[5], which focuses on the societal consequences of the rapid development of artificial intelligence (AI) and automated decision-making (ADM) in a wide variety of societal areas. Human(e) AI is built on inter-faculty collaboration between the faculties of science, law, humanities and social and behavioral sciences.

3. GRAND AI RESEARCH CHALLENGES

AI SYSTEMS & HUMANS

♦ ***ALLAI recommends to add a research question on the impact of AI on work***

o *Additional RQ: How can we make sure that AI in the workplace is usable and that the worker still has sufficient autonomy and control, fulfillment and job satisfaction?*

AI will not only affect the quantity of available work but also the nature of existing work. AI systems offer more and more opportunities to track and monitor workers, raising concerns over autonomy and privacy. Work is now often determined and distributed by algorithms without human intervention, which influences the nature of the work as well as working conditions. There is also the risk of a drop in the quality of jobs and the loss of important skills through the use of AI systems.

AI can have major benefits when used for hazardous, heavy, exhausting, dirty, unpleasant, repetitive or boring work, but also data processing and analysis, planning and prediction can increasingly be performed by AI systems. AI systems are increasingly being used to monitor and monitor workers, thereby compromising autonomy and privacy. Work is increasingly defined and distributed by algorithms, without human intervention, which influences the nature of the work and working conditions. And there is a risk of a loss of work and a loss of skills.

[5] https://humane-ai.nl/

It is important to focus not only on what AI is capable of doing, but also on what people are capable of doing (creativity, empathy, cooperation) and what we want people to keep doing, and to look for opportunities to enable people and machines to work together better.[6]

Augmented intelligence (complementarity), whereby human and machine work together and support each other, is the most interesting application of AI since its involves human with machine, as opposed to human instead of machine.[7] However, co-creation is of major importance: workers must be involved in developing these kinds of complementary AI systems, in order to ensure that the systems are useable and that the worker still has sufficient autonomy and control (human-in-command), fulfilment and job satisfaction.

AI SYSTEMS & SOCIETY

♦ ***ALLAI recommends to add a research question on 'human agency and oversight'***

o *Additional RQ: How can we ensure human agency and oversight in the development, deployment and use of AI?*

AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and foster fundamental rights, and allow for human oversight.

Human agency means that users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system. AI systems should support individuals in making better, more informed choices in accordance with their goals. AI systems can sometimes be deployed to shape and influence human behavior through mechanisms that may be difficult to detect, since they may harness sub-conscious processes, including various forms of unfair manipulation, deception, herding and conditioning, all of which may threaten individual autonomy.

Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such a s a human-in-the- loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach.

HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation.

[6] EESC INT/806, AI & Society, 2017, Rapporteur: Catelijne Muller.

Human-in-command (HIC) refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system. [8]

Moreover, it must be ensured that public enforcers have the ability to exercise oversight in line with their mandate. Oversight mechanisms can be required in varying degrees to support other safety and control measures, depending on the AI system's application area and potential risk. All other things being equal, the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required.

- *Additional RQ How can we make sure that AI development occurs in the most environmentally friendly way possible?*

AI systems promise to help tackling some of the most pressing societal concerns, yet it must be ensured that this occurs in the most environmentally friendly way possible. The system's development, deployment and use process, as well as its entire supply chain, should be assessed in this regard, e.g. via a critical examination of the resource usage and energy consumption during training, opting for less harmful choices. Measures securing the environmental friendliness of AI systems' entire supply chain should be encouraged.

<u>CROSS-CUTTING CONCERNS</u>

- ***ALLAI recommends to also add 'human agency and oversight' (see above) as a cross-cutting concern***

- ***ALLAI recommends to add 'technical robustness, accuracy and safety'[9] as a cross-cutting concern***

Technical robustness requires that AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimizing unintentional and unexpected harm, and preventing unacceptable harm. This should also apply to potential changes in their operating environment or the presence of other agents (human and artificial) that may interact with the system in an adversarial manner. In addition, the physical and mental integrity of humans should be ensured.

AI systems, like all software systems, should be protected against vulnerabilities that can allow them to be exploited by adversaries, e.g. hacking. Attacks may target the data (data poisoning), the model (model leakage) or the underlying infrastructure, both software and hardware. AI systems should have safeguards and fall back plans in case of problems or attacks.

[8] VPRO Tegenlicht, Mens in de Machine, met Catelijne Muller over *human-in-command*, 42:20'; EESC INT/806, AI & Society, 2017, Rapporteur: Catelijne Muller; Policy and Investment Recommendations for Trustworthy AI (AI HLEG)

[9] Policy and Investment Recommendations for Trustworthy AI, High Level Expert Group on AI, 2019.

It must be ensured that the system will do what it is supposed to do without harming living beings or the environment. This includes the minimization of unintended consequences and errors. In addition, processes to clarify and assess potential risks associated with the use of AI systems, across various application areas, should be established. AI systems should be accurate, i.e. contain the ability to make correct judgements, for example to correctly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models. An explicit and well-formed development and evaluation process can support, mitigate and correct unintended risks from inaccurate predictions.

It is critical that the results of AI systems are reproducible, as well as reliable. A reliable AI system is one that works properly with a range of inputs and in a range of situations. This is needed to scrutinize an AI system and to prevent unintended harms. Reproducibility describes whether an AI experiment exhibits the same behavior when repeated under the same conditions.

5. IMPACT OF AI IN APPLICATION DOMAINS

- ***ALLAI recommends focusing research on sectors in which the Netherlands is already strong***

Apart from fundamental research in AI and the ethical, legal and societal implications of AI, research should focus on sectors in which the Netherlands is already strong, including 'Water and Maritime', 'Agri and Food', 'Life Sciences & Health' and 'High-tech Systems and Materials'. AI can provide economic, social and environmental benefits in these sectors.

ALLAI.

ALLAI refers to Stichting ALLAI Nederland, een foundation under Dutch Law. No entity or person related to ALLAI is liable or eny direct or indirect loss or damage as a result of an entity or person that relying on this communication.

ALLAI is Stichting ALLAI Nederland, een stichting naar Nederlands recht Geen enkele entiteit of persoon gelieerd aan ALLAI is verantwoordelijk of aansprakelijk voor enig direct of indirect verlies of enige schade geleden door een entiteit of persoon die op deze communicatie vertrouwt.

Herengracht 247
1016 BH Amsterdam
welkom@allai.nl
www.allai.nl