RAAIT Workshop ECAI 2023

summary

At the European Conference on Artificial Intelligence (ECAI) 2023, held in Poland, Krakow, the first workshop on Responsible Applied Artificial InTelligence (RAAIT) took place. Below you will find a summary of the contributions of our speaker and participating visitors and the insights gained through the day's valuable discussions. We invite you to read the full papers of contributing authors through the bibliography at the end of this document.

Proceedings will be published including many of the underlying papers of presentations held at our workshop.

Keynote

Our keynote speaker, **Emma Beauxis-Aussalet**, held a presentation on the prerequisites and challenges for making AI work ethically in practice. She showcased that even if the math underlaying your AI model is correct, if you can't communicate it, or if it's not comprehensive, it's likely to not work. She made the comparison to how sick cows in the past used to be held up by a pulley system to produce cheap - but bad - milk, and how currently eXplainable AI (XAI) is also used in certain organisations as a pulley system to keep a bad AI program up and running with bad results as a consequence.

Presentations

We were fortunate to have had many engaging presentations by other contributing researchers, and valuable follow-up dialogues.

András Strausz and **Ziyao Shang** showed us their implementation for a human-in-the-loop framework, built on a flexible web interface, to support researchers in their exploration of bias present in their visual datasets.

Laura de Groot showed us their physical model for visualising Machine Learning (ML) tradeoffs, supporting the integration of (non-)expert stakeholders into AI development practice.

Stefan Leijnen showed us the intriguing balance between showing users the complexity of a system and its effect on trust, where showcasing the system's complexity can lead both to an increase in warranted trust but also in unwarranted distrust.

Steven Vethman and **Cor Veenman** showcased their method to allow test labs to perform pilot studies without adversely impacting (un)known participating individuals, meanwhile increasing the amount of actionable insight gained through the study.

Max Knobbout introduced an adversarial method to achieve data fairness by erasing sensitive variables. Two artificial agents compete; the actor hides the sensitive variable using an encoder, while the adversary tries to guess it from the encoded data. This process continues until the adversary can't accurately guess the sensitive variable, ensuring a fair dataset representation.

Jacintha Walters presented research on compliance with the EU AI Act. Through a series of interviews with small and medium companies, she outlined the most important topics they will have to deal with.

Felix Friedrich showed how to explicitly steer a generative image model away from certain unintended biases it may have through a low-threshold approach, in contrast to negative prompting where unintended biases will continue to show up.

Martin van den Berg shared insights from interviews at a financial firm, focusing on a use case of fraudulent insurance claims. Participants discussed the process, ethical assessments, and practices for managing risks and negative impacts for stakeholders or beyond.

Coert van Gemeren outlined their research into the practical applicability of AI Impact Assessments in the media sector. Through interviews, they investigated what parts of the AI Impact Assessment created by the Netherlands AI Coalition could benefit the media sector, and how such an instrument could be tweaked in order for it to be useful and suitable for adoption in this sector.

Highlights / trends/ observations

During the presentations, we mapped the topics of the presentations on prepared boards, noticing highlights and trends.There's a complexity in ensuring ethical outcomes of AI systems and design, where even completely ethical decision processes can lead to unethical outcomes, and but a single error can chain down to many down the line. As such, it's important to design for failure, making sure the systems 'break properly' and that we're already prepared for what follows.

There's furthermore a necessity to integrate stakeholders properly (not as a means to perform 'participation washing') as early as possible into the AI design process, ensuring they are not negatively impacted by our tests for validation, and letting them steer us towards potential sources of bias in our systems.

Participants of the workshop would love to see more open engagement between design science and engineering as a two-way street, rather than a one-way street from the latter to the former. More room should be made at conferences for user studies and validation, as it seems to be undervalued in computer science at the moment.

Challenges from breakout

During our break-out session, the group was subdivided in 4, tasked with deciding upon the top 3 challenges to making Responsible AI in practice possible, focusing on different sub-topics: values, organizational context, AI-application and methods.   These four sub-topics are based on the RAAIT's research agenda, which you can read more about here. The results are listed here below:

Values

1. Balancing focusing on specified values and facilitating an open discussion.
2. Values often stay vague or broad. One challenge is the difficulty in making values concrete and internalise these throughout the entire organisation.
3. Including societal and ecological issues as part of the values we build Responsible AI on.

Organizational context

1. The effect of the AI transition on the job market: Will jobs become obsolete, should jobs be protected?
2. Resilience versus dependence on AI. There currently is a big reliance on organisations such as Huggingface and OpenAI, but what if they disappear? How are we going to make assurances for these scenarios, and become more resilient in face of change?
3. How do we reconcile ethics and capitalism? Shareholder value is the biggest driver for companies. How do you align ethical values with shareholder interests.

AI Applications

1. Lack of economic incentive – or clarity of such – to pursue ethical development of AI for companies.
2. There is no single solution that works for every organisation, making the bar for entry high
3. Involving all stakeholders throughout the entire development pipeline is often seen as a threat/liability/chore to the 'proper' functioning of the development process.

Methods

1. Development metrics for Responsible AI
2. Application of metrics for Responsible AI
3. There is a difficulty in cooperating with engineering due to current interactions being a one-way street of engineering to design studies.

Conclusion

The great contributions of participants, the engaging discussions and through it our own insights, underline the necessity for an interdisciplinary, stakeholder-engaged, ethically conscious approach to AI development and deployment. We hope to do our part in this through our work at RAAIT.

It is our ambition to grow RAAIT to a successful yearly event, and are looking for people like You to help contribute to its impact, whether big or small, as in the end we have to work together to ensure that Responsible AI research actually gets put into practice.

If you're so inclined, see the notes below for points of contact:

- Help organise the next edition?
  Mail: info@raait-ecai-2023.com
- Want a heads-up about the next edition, or for when its call for papers opens?
  link to mail list
- Interested in the larger RAAIT program (it's not just a yearly workshop!)
  Website: raait.nl

And of course, feel free to contact us in general with questions and/or remarks at info@raait-ecai-2023.com.

Thank you so much for your interest in our workshop, and let's actually make some impact on responsible AI in practice!

Contributions

Max Knobbout, *ALFR++: A novel algorithm for Learning Adversarial Fair Representations*

Laura de Groot, *The Machine Vision Game: Making Machine Vision Development Trade-Offs Tangible*

Senthuran Kalananthan, Alexander Kichutkin, Ziyao Shang, András Strausz, Javier Sanguino and Menna Elassady, *MindSet: A Data-Debiasing Interface using a Visual Human-in-the-Loop Workflow*

Felix Friedrich, Manuel Brack, Patrick Schramowski and Kristian Kersting, *Mitigating Inappropriateness in Image Generation: Can there be Value in Reflecting the World's Ugliness?*

Floor Schukking, Levi Verhoef, Tina Mioch, Coert van Gemeren and Huib Aldewereld, *Improving Adoption of AI Impact Assessment in the Media Sector*

Danielle Sent, Tina Wünn and Linda W.P. Peute, *Trust in Artificial Intelligence: Exploring the Influence of Model Presentation and Model Interaction on Trust in a Medical Setting*

Martin van den Berg, Julie Gerlings and Jenia Kim, *Empirical Research on Ensuring Ethical AI in Fraud Detection of Insurance Claims: A Field Study of Dutch Insurers*

Steven Vethman, Marianne Schaaphok, Marissa Hoekstra and Cor Veenman, *Random Sample as a Pre-Pilot Evaluation of Benefits and Risks for AI in Public Sector*

Jacintha Walters, Diptish Dey, Debarati Bhaumik, Sophie Horsman, *Complying with the EU AI Act, On which areas should organizations focus when considering compliance with the AIA?*