Field of Action
#1 Data Governance: Quality

Why is this topic challenging?

AI is about learning from data. And a huge focus has been on the models, training methods, and so on. With great success!

But at the same time, data has always been seen as "given", and it's a bit of a chicken–egg problem to quantify how good your data are. Since, for that, you need to have a way of representing it in a machine-learning model, and…

Still, we need to ensure that the data that is used for a particular model is representative of the underlying problem. For instance, you don't use a camera-based skin cancer detection method trained on fish for people living in Lapland.

And apart from obvious (?) data privacy issues, data governance for AI must also address ethical considerations, such as fairness, transparency, and accountability.

Why a collaboration of research & commercial partners?

The topic is challenging from an academic perspective; currently, no widely accepted data documentation methods are established which can be directly used. It is very likely that scientific publications come out of this collaboration. Conversely, Industry collaboration can bring practical perspectives, and real-world data cq. use cases, and those real use cases can lead to applicable solutions.

Description of the partners needed

At least two partners, one academic, one commercial. Technical background in machine learning is highly recommended for at least one of the partners.

Pilot example

YZ Company is looking to establish a data governance framework to comply with one of the high-risk requirements of the AI Act. They have a vast amount of historical data. However, they recognize the need to ensure that their AI models are built on high-quality, unbiased data and comply with regulatory requirements.

References to scientific publications

McMillan-Major, A., Bender, E. M., & Friedman, B. (2023). Data statements: From technical concept to community practice. ACM Journal on Responsible Computing.

Pansara, R. (2023). Cultivating Data Quality to Strategies, Challenges, and Impact on Decision-Making. International Journal of Management Education for Sustainable Development, 6(6), 24-33.

Field of Action
#1 Data Governance: Documentation

Why is this topic challenging?
AI is about learning from data. And a huge focus has been on the models, training methods, and so on. With great success!

But at the same time, data has always been seen as "given", and it's a bit of a chicken–egg problem to quantify how good your data are. Since, for that, you need to have a way of representing it in a machine-learning model, and…

Still, we need to ensure that the data that is used for a particular model is representative of the underlying problem. This is addressed by the Field of Action #1 Data Governance: *Quality*.

Here, we look at tools to support that: how can we document our data sets, and how can we document the improvement and adaptation of data to fit quality criteria in the Field of Action #1 Data Governance: *Quality*. How can we prepare our data for re-use, …

In short, #1 Data Governance: *Documentation* is about developing and evaluating data documentation methodologies.

Why a collaboration of research & commercial partners?
The topic is challenging from an academic perspective; currently, no widely accepted data documentation methods are established which can be directly used. It is very likely that scientific publications come out of this collaboration. Conversely, Industry collaboration can bring practical perspectives, and real-world data cq. use cases, and those real use cases can lead to applicable solutions.

Description of the partners needed
At least two partners, one academic, one commercial. Technical background in machine learning is highly recommended for at least one of the partners.

Pilot example
The company has data sets that need to be documented. Existing methods can be evaluated, expanded, and possibly published.

References to scientific publications

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

McMillan-Major, A., Bender, E. M., & Friedman, B. (2023). Data statements: From technical concept to community practice. ACM Journal on Responsible Computing.

Roman, A. C., Vaughan, J. W., See, V., Ballard, S., Schifano, N., Torres, J., ... & Ferres, J. M. L. (2023). Open Datasheets: Machine-readable Documentation for Open Datasets and Responsible AI Assessments. arXiv preprint arXiv:2312.06153.

Field of Action
#2 Human Oversight

Why is this topic challenging?

According to the AI Act, Article 14: “High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which the AI system is in use.” And it can go so far as to define a “stop button”.

But how can this be done? In which cases can we have humans oversee AI systems, which are much faster and typically black-box? How does this relate to explainability of AI systems? What expertise, what tools are needed? What is the economic impact? Will this lead to offshoring; what is the social impact? Moreover, how can we make sure that people operating AI understand their limitations and do not over-rely on such systems?

Why a collaboration of research & commercial partners?

This topic has a strong academic component, since there is little methodological ground covered. Still, use-case based collaboration with non-academic partners is key, not just to validate methodologies in real applications but also to learn which end-user challenges exist.

Description of the partners needed

IT giants have already implemented various levels of automated and human oversight of social media posts, from which we have learned that this topic has a strong sociological component as well as a macroeconomic perspective. At the same time, understanding of the life cycle model of AI systems is needed.

Pilot example

A kitchen gas burner has an AI-based safety component, used to detect leaks with a novel gas sensor. How do we implement human oversight? Since remote human oversight at scale will probably not work here, this falls in the category of “measures identified by the provider before placing the high-risk AI system on the market or putting it into service and that are appropriate to be implemented by the user.” How?

References to scientific publications

Enqvist, L. (2023). 'Human oversight in the EU artificial intelligence act: what, when and by whom?. Law, Innovation and Technology, 15(2), 508-535.

Pavlidis, G. (2024). Unlocking the black box: analysing the EU artificial intelligence act's framework for explainability in AI. Law, Innovation and Technology, 1-16.

Raji, I. D., & Yang, J. (2019). About ml: Annotation and benchmarking on understanding and transparency of machine learning lifecycles. arXiv preprint arXiv:1912.06166.

Field of Action
#3 Accuracy, Robustness, Cybersecurity

Why is this topic challenging?

This actually consists of three topics which are closely related. The AI Act says, “High-risk AI systems should perform consistently throughout their lifecycle and meet an appropriate level of accuracy, robustness and cybersecurity, in the light of their intended purpose and in accordance with the generally acknowledged state of the art.” Vague enough.

The relation between the three topics comes from the fact that they are all describing properties of AI methods that increase their trust. *Accuracy* means, the errors in replies by the AI system are within predefined limits.

Within this topic, several technical questions must be answered. First, how are these limits determined? Second, what does it mean to stay between those limits, depending on the nature of outliers? Then, for instance, how does the interconnected nature of AI systems influence these answers? The range of potential topics is wide.

Why a collaboration of research & commercial partners?

The topic involves the transfer of technical knowledge from academic partners to commercial partners. Possibly, a commercial partner with strong technical knowledge w.r.t. machine learning might work autonomously but risks not having a wide enough scope on technical solutions.

Description of the partners needed

The topic requires good knowledge of the machine-learning landscape, therefore strong technical expertise is recommended, especially at the academic side. A strong industrial partner is recommended to ensure rigorous testing and verification of the suggested approaches.

Pilot example

A possible pilot can focus on outlier detection, e.g. in “predictive maintenance” and the kind. A more interesting pilot, albeit perhaps more scientific, may be about accuracy in large-language models by looking at, e.g., factuality.

References to scientific publications

Ashoori, M., & Weisz, J. D. (2019). In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. arXiv preprint arXiv:1912.02675.

Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It's complicated: The relationship between user trust, model accuracy and explanations in ai. ACM Transactions on Computer-Human Interaction (TOCHI), 29(4), 1-33.

van der Veer, S. N., Riste, L., Cheraghi-Sohi, S., Phipps, D. L., Tully, M. P., Bozentko, K., ... & Peek, N. (2021). Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. Journal of the American Medical Informatics Association, 28(10), 2128-2138.

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019, May). Understanding the effect of accuracy on trust in machine learning models. In Proceedings of the 2019 chi conference on human factors in computing systems (pp. 1-12).

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020, January). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 295-305).

Field of Action

#4 Accuracy, Robustness, Cybersecurity

Why is this topic challenging?

This actually consists of three topics which are closely related. The AI Act says, “High-risk AI systems should perform consistently throughout their lifecycle and meet an appropriate level of accuracy, robustness and cybersecurity, in the light of their intended purpose and in accordance with the generally acknowledged state of the art.” Vague enough.

The relation between the three topics comes from the fact that they are all describing properties of AI methods that increase their trust. *Robustness* means, the system can handle outliers well.

Here it is not only challenging to ensure that the systems are robust enough, but also to develop methods to evaluate that at all. After all, complete testing is not possible due to the high-dimensional nature of the input data.

Why a collaboration of research & commercial partners?

The topic involves the transfer of technical knowledge from academic partners to commercial partners. Possibly, a commercial partner with strong technical knowledge w.r.t. machine learning might work autonomously but risks not having a wide enough scope on technical solutions.

Description of the partners needed

The topic requires good knowledge of the machine-learning landscape, therefore strong technical expertise is recommended, especially at the academic side. A strong industrial partner is recommended to ensure rigorous testing and verification of the suggested approaches.

Pilot example

There are several approaches towards robustness, including adversarial testing, stress testing, noise imputation, checking robustness against shifts of data distributions, real-world testing (!), analysing accuracy / precision / recall. This can be done for almost any system.

References to scientific publications

Chang, H., Nguyen, T. D., Murakonda, S. K., Kazemi, E., & Shokri, R. (2020). On adversarial bias and the robustness of fair machine learning. arXiv preprint arXiv:2006.08669.

Lee, J. G., Roh, Y., Song, H., & Whang, S. E. (2021, August). Machine learning robustness, fairness, and their convergence. In Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining (pp. 4046-4047).

Hancox-Li, L. (2020, January). Robustness in machine learning explanations: does it matter?. In Proceedings of the 2020 conference on fairness, accountability, and transparency(pp. 640-647).

Rauber, J., Brendel, W., & Bethge, M. (2017). Foolbox: A python toolbox to benchmark the robustness of machine learning models. arXiv preprint arXiv:1707.04131.

Field of Action

#5 Accuracy, Robustness, Cybersecurity

Why is this topic challenging?

This actually consists of three topics which are closely related. The AI Act says, “High-risk AI systems should perform consistently throughout their lifecycle and meet an appropriate level of accuracy, robustness and cybersecurity, in the light of their intended purpose and in accordance with the generally acknowledged state of the art.” Vague enough.

The relation between the three topics comes from the fact that they are all describing properties of AI methods that increase their trust. *Cybersecurity* means, the system handles external cyberattacks gracefully. This includes data poisoning, adversarial attacks, inclusion of faulty pre-trained components, etc.

Cybersecurity in AI systems is a relatively new topic, but many subtopics have a good history in the scientific community.

Why a collaboration of research & commercial partners?

The topic involves the transfer of technical knowledge from academic partners to commercial partners. Possibly, a commercial partner with strong technical knowledge w.r.t. machine learning might work autonomously but risks not having a wide enough scope on technical solutions.

Description of the partners needed

The topic requires good knowledge of the machine-learning landscape, therefore strong technical expertise is recommended, especially at the academic side. A strong industrial partner is recommended to ensure rigorous testing and verification of the suggested approaches.

Pilot example

As for robustness, cybersecurity evaluation should be performed for any system that receives outside data. Standard cybersecurity evaluation methods can be used.

References to scientific publications

Breier, J., Baldwin, A., Balinsky, H., & Liu, Y. (2020). Risk Management Framework for Machine Learning Security. arXiv preprint arXiv:2012.04884.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228.

Lee, I. (2021). Cybersecurity: Risk management framework and investment cost analysis. Business Horizons, 64(5), 659-671.

Shah, V. (2021). Machine Learning Algorithms for Cybersecurity: Detecting and Preventing Threats. Revista Espanola de Documentacion Cientifica, 15(4), 42-66.

Field of Action
#6 Risk Management (Safety)

Why is this topic challenging?

A risk-management system is an organisational process that places assessing and mitigating risks at the heart of the entire design and development process. This is therefore an encompassing approach enabling the other topics that need to be done – within company processes. The risk management system will help AI providers determine, for instance, what an "effective" level of human oversight is for their product or what statistical properties are "appropriate" for their data given the risks.

Implementing a risk management system requires product manufacturers to establish and document responsibilities, risk management steps, and methods used in assessing and evaluating risks.

Why a collaboration of research & commercial partners?

Given the capacity, this topic can be addressed by companies themselves. However, knowledge on AI life cycle models is a clear plus, and that is not established, rather a matter of ongoing research and development. Then, also, the capacity needed is rather broad, requiring knowledge on business process methodologies as well as, as mentioned, life cycle models.

The corresponding knowledge may, in part, be provided by consulting companies.

Description of the partners needed

From the academic side, partners with a background in process optimisation, business management, etc. are ideally suited. Yet also, knowledge of AI life cycle models and corresponding governance frameworks is required, either in that or in an additional partner. From a commercial perspective, the ideal partner would be an established SME with a clear but adaptable process landscape. Multiple of these would be ideal.

Pilot example

In a pilot, the participating companies would serve as a testbed for implementing a risk management system for AI. Using scientific risk management concepts, life cycle models, etc., an implementation can be realised. The developed framework undergoes rigorous validation through simulations, case studies, and real-world scenarios. Feedback loops are established to gather input from diverse stakeholders, including end-users, regulators, and advocacy groups, ensuring that the framework reflects a wide range of perspectives.

Feedback from the participating companies would involve assessing the cost (effort, time, …) for implementation as well as the impact on the AI development process. Upon finalizing the risk assessment framework, comprehensive documentation is prepared to facilitate its adoption and implementation within the companies' AI development lifecycle.

References to scientific publications

AI, Nist. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0).

Alzubaidi, L., Al-Sabaawi, A., Bai, J., Dukhan, A., Alkenani, A. H., Al-Asadi, A., ... & Gu, Y. (2023). Towards risk-free trustworthy artificial intelligence: Significance and requirements. International Journal of Intelligent Systems, 2023.

Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2024). AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act. Digital Society, 3(1), 1-29.

Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2023). How to evaluate the risks of artificial intelligence: a proportionality-based, risk model for the AI Act. Risk Model for the AI Act (May 31, 2023).

Field of Action
#7 Post-Market Monitoring

Why is this topic challenging?
Due to the complexity of AI systems it is unpredictable how an algorithm would perform on the market. This can lead to unforeseen risks that must be effectively monitored and mitigated. Once an AI system is in use, one often sees that real-world data differs significantly from the training data, leading to errors or worse.

Continuous monitoring ensures that AI systems remain safe and perform reliably as intended, even as they encounter new data or conditions post-deployment. This is crucial for maintaining user trust and for the safety of AI applications in critical areas like healthcare, finance, or transportation.

Why a collaboration of research & commercial partners?
The tools available for post-market monitoring are sparse when it comes to AI systems. There is a strong research element involved in this work. At the same time, the validity of that work only comes with “real” deployment of AI systems – ideally in sandboxes, or for low-risk systems.

We do not expect a one-fit-for-all solution; multiple of these can run in parallel with different end users (applicators).

Description of the partners needed
This highly complex topic requires both technical knowledge and ethical expertise. An ideal setup would be a triumvirate of two scientific/technical partners, and an end-user. The end-user would ideally be a startup, or possibly an SME, wanting to put a system on the market.

Pilot example
The pilot should begin with the assessment of the effectiveness and limitations of the company's current post-market monitoring methods for AI systems. The team has to collaboratively identify relevant performance metrics and indicators for monitoring AI systems in real-world settings. The goal of the pilot is to develop an enhanced framework for post-market monitoring, integrating advanced analytics, machine learning algorithms, and real-time data processing capabilities that monitor the performance of a real-world AI system in the market.

References to scientific publications

Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., ... & Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development, 63(4/5), 6-1.

Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., de Prado, M. L., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Information Fusion, 99, 101896.

Zicari, R. V., Brodersen, J., Brusseau, J., Düdder, B., Eichhorn, T., Ivanov, T., ... & Westerlund, M. (2021). Z-Inspection®: a process to assess trustworthy AI. IEEE Transactions on Technology and Society, 2(2), 83-97.