Safety in AI

Holger H. Hoos Marie Anastacio Jakob Bossek

21 October 2022

1 Seminar description

Prof. Holger H. Hoos & M.Sc. Marie Anastacio & Dr. Jakob Bossek
Chair for AI Methodology (Informatik 14)
Website: `http://www.aim.rwth-aachen.de/`

While in the past, performance has been the main focus of much work in AI, aspects of safety are increasingly recognised as similarly significant. This block seminar course on AI Safety will be held in English, towards the end of the 2022/2023 winter semester. Enrolment is restricted to 20 Master or senior Bachelor students, preferably with a background in AI (including methods from machine learning, optimisation, planning and scheduling, multi-agent systems and other areas of AI). Students will work in groups of two on a range of topics from AI Safety, which spans the robustness and verification of AI methods, including - but not restricted to - neural networks; the explainability and interpretability of the results obtained from AI algorithms; as well as bias and privacy issues in machine learning. Each group will be assigned recently published work from the research literature, which will serve as the starting point for an in-depth investigation of a specific topic; the results of this investigation will be presented in class and compiled into a report.

2 Seminar procedure

In an introductory *kick-off meeting* we will present our ideas on the seminar procedure. Students will be divided into groups of two by us using a semi-random process aimed at ensuring diversity and complementarity of experience within the groups. Each group will be assigned a recent publication from the field, which serves as a starting point into the respective topic. The groups dive into the topic by performing literature search and compile a survey-like report giving an overview of the respective field. The results are presented in an oral presentation.

- The seminar will take place as a block-seminar in February or March 2023.

- 30 minutes talk (each student must contribute equally) plus additional 30 minutes of in-depth discussion.
- Seminar report: 20 pages max, using the LaTeX template provided by us, including references, figures etc. A statement outlining the contributions of each team member is mandatory and will be used as one basis for assessment.

3 Oral Presentation

- Use the provided LaTeX-template (see website) for the presentation slides (we do not allow Power-Point presentations).
- To not lose yourself in unimportant details too much.
- Keep the time limit (30min; a little less is OK, a little more is *not* OK).
- Each group member should participate equally.

4 Report

- Note that the paper assigned to your group is *not necessarily the most relevant*. We expect you to take it as a 'first clue', deep-dive into the literature and compile the most relevant aspects. Discovering and deciding which papers are important is part of your work. Note also that not everything has to be covered in full detail. It is up to the group to decide which papers and content is most relevant.
- Use the provided LaTeX-template (see website) for the final report and submit in PDF-format (we do not accept MS-Word reports).
- Stick to the page limit: 20 pages using the prescribed format, including references, figures etc.
- A report contains introduction, conclusion and bibliography among other sections.
- Additional material on how to write good papers will be made available no later than June 15.

5 Criteria for successful completion

- Preparation of a seminar report in LaTeX (max. 20 pages, using the prescribed format, PDF)
- 30 minute presentation + 30 minutes discussion
- Meeting all deadlines

- Attendance of all mandatory meetings
- **Grading**: 60% report, 30% presentation incl. answers to questions and 10% participation in discussions on other presentations.

6 Important Dates

- Kickoff meeting: 21 October 2022
- Progress update (via e-mail, bullet points are OK, but do give us some details): **18 November 2022 6pm CEST** (hard deadline!)
- Final report due (PDF via e-mail): **27 January 2023 6pm CEST** (hard deadline!)
- Block seminar: February or March 2023 (tba)

7 Groups and topics

Group assignment performed by random permutation while making sure that no two Bachelor students are assigned the same group. I. e., each group has at least one Master student.

RV-1 Konstantin Geisler, Rene Heinz-Peter Evertz
Topic: Adversarial attacks
Christian Szegedy et al. "Intriguing properties of neural networks". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2014

RV-2 Luk Jonas Fuchs, Nick Valentin Kocher
Topic: Formal verification
Vincent Tjeng, Kai Yuanqing Xiao, and Russ Tedrake. "Evaluating Robustness of Neural Networks with Mixed Integer Programming". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019

RV-3 Kiana Mishelle Adamik, Nicolas Schumann
Topic: Heuristic verification
Nicholas Carlini and David A. Wagner. "Towards Evaluating the Robustness of Neural Networks". In: *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*. IEEE Computer Society, 2017, pp. 39–57. DOI: 10.1109/SP.2017.49

EI-1 Nhu-Yen Nguyen, Marius Andre Jean-Michel Peterfalvi
Topic: Outcome explanation (global)
Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model

Predictions". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. ed. by Isabelle Guyon et al. 2017, pp. 4765–4774

EI-2 Torge Schöwing, Jan Brinkmann
Topic: Outcome explanation (local)
Bolei Zhou et al. "Learning Deep Features for Discriminative Localization". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2921–2929. DOI: 10.1109/CVPR.2016.319

EI-3 Ozan Ege Sap, Peter Benjamin Schwarz
Topic: Inspection
Julia Moosbauer et al. "Explaining Hyperparameter Optimization via Partial Dependence Plots". In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*. Ed. by Marc'Aurelio Ranzato et al. 2021, pp. 2280–2291

EI-4 Dobromir Iordanov Panayotov, Katharina Kösseler
Topic: Counterfactual explanations
Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations". In: *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*. Ed. by Mireille Hildebrandt et al. ACM, 2020, pp. 607–617. DOI: 10.1145/3351095.3372850

P-1 Yitong Guo, Niklas Fücker
Topic: Model inversion
Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-16, 2015*. Ed. by Indrajit Ray, Ninghui Li, and Christopher Kruegel. ACM, 2015, pp. 1322–1333. DOI: 10.1145/2810103.2813677

P-2 Phillip Ahlers, Mohamed-Wassim Deghdagh
Topic: Inference
Karan Ganju et al. "Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations". In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018*. Ed. by David Lie et al. ACM, 2018, pp. 619–633. DOI: 10.1145/3243734.3243834

B-1 Archit Dhama, Marco Bischoff
Topic: Mitigating bias via sampling strategies
F Kamiran and TGK Calders. "Classification with no discrimination by preferential sampling". In: *Informal proceedings of the 19th Annual Machine Learning*

Conference of Belgium and The Netherlands (Benelearn'10, Leuven, Belgium, May 27-28, 2010). 2010, pp. 1–6

B-2 Xiyang Yang, Julian Treiber

Topic: Mitigating bias via data alteration

Hao Wang, Berk Ustun, and Flávio P. Calmon. "Repairing without Retraining: Avoiding Disparate Impact with Counterfactual Distributions". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 6618–6627