Report from Dagstuhl Seminar 20382

Interactive Visualization for Fostering Trust in AI

Edited by

Daniela Oelke¹, Daniel A. Keim², Polo Chau³, and Alex Endert⁴

- 1 Hochschule Offenburg, DE, daniela.oelke@hs-offenburg.de
- 2 Universität Konstanz, DE, keim@uni-konstanz.de
- 3 Georgia Tech, US, polo@gatech.edu
- 4 Georgia Tech, US, endert@gatech.edu

— Abstract -

Artificial intelligence (AI), and in particular machine learning algorithms, are of increasing importance in many application areas but interpretability and understandability as well as responsibility, accountability, and fairness of the algorithms' results, all crucial for increasing the humans' trust into the systems, are still largely missing. Big industrial players, including Google, Microsoft, and Apple, have become aware of this gap and recently published their own guidelines for the use of AI in order to promote fairness, trust, interpretability, and other goals. Interactive visualization is one of the technologies that may help to increase trust in AI systems. During the seminar, we discussed the requirements for trustworthy AI systems as well as the technological possibilities provided by interactive visualizations to increase human trust in AI.

Seminar September 13–16, 2020 – http://www.dagstuhl.de/20382

2012 ACM Subject Classification Computing methodologies \rightarrow Artificial intelligence, Humancentered computing \rightarrow Visualization, Computing methodologies \rightarrow Machine learning

Keywords and phrases accountability, artificial intelligence, explainability, fairness, interactive visualization, machine learning, responsibility, trust, understandability

Digital Object Identifier 10.4230/DagRep.10.4.37

1 Executive Summary

Daniela Oelke Daniel Keim Polo Chau Alex Endert

License

Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
Creative Commons BY 3.0 Unported license

License
License
Creative Commons BY 3.0 Unported license

License
License
Creative Commons BY 3.0 Unported license

License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License
License

Artificial Intelligence (AI) and other computational processes continue to influence decisions across a wide range of applications including healthcare decisions, vehicle navigation, data science, and others. This Dagstuhl seminar reflected on some of the challenges inherent in the goal of increasing the interpretability of these systems, and when applicable, increase the trust people put into them to make decisions. The seminar participants discussed the complexity of trust itself, and how the concept is multi-faceted, and likely outside of researchers in technology and computer science to fully define. We discussed an inter-disciplinary research agenda, as well as a manifesto that should help frame this direction going forward.

2 Table of Contents

Executive Summary Daniela Oelke, Daniel Keim, Polo Chau, and Alex Endert	37
Overview of Talks	
Mixed-Initiative Topic Model Refinement through Visual Analytics Mennatallah El-Assady	39
Trustworthy AI – A Medical Perspective Jörn Kohlhammer	39
Visual Analytics for Large-Scale ML Systems Minsuk Khang 4	40
Autonomy and trust in spaceflight operationsScott Davidoff4	40
ExplainExplore – Experimenting with ExplanationsJarke J. van Wijk	11
Participants	12

3 Overview of Talks

3.1 Mixed-Initiative Topic Model Refinement through Visual Analytics

Mennatallah El-Assady (Universität Konstanz, DE)

License	Creative Commons BY 3.0 Unported license
	© Mennatallah El-Assady
Joint work of	Mennatallah El-Assady, Rita Sevastjanova, Fabian Sperrle, Rebecca Kehlbeck, Daniel A. Keim,
	Christopher Collins, Oliver Deussen
Main reference	Mennatallah El-Assady, Rita Sevastjanova, Fabian Sperrle, Daniel A. Keim, Christopher Collins:
	"Progressive Learning of Topic Modeling Parameters: A Visual Analytics Framework", IEEE Trans.
	Vis. Comput. Graph., Vol. 24(1), pp. 382–391, 2018.
URL	https://doi.org/10.1109/TVCG.2017.2745080
Main reference	Mennatallah El-Assady, Fabian Sperrle, Oliver Deussen, Daniel A. Keim, Christopher Collins:
	"Visual Analytics for Topic Model Optimization based on User-Steerable Speculative Execution",
	IEEE Trans. Vis. Comput. Graph., Vol. 25(1), pp. 374–384, 2019.
URL	http://dx.doi.org/10.1109/TVCG.2018.2864769
Main reference	Mennatallah El-Assady, Rebecca Kehlbeck, Christopher Collins, Daniel A. Keim, Oliver Deussen:
	"Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections",
	IEEE Trans Vis Comput. Graph. Vol 26(1), pp 1001–1011, 2020

URL https://doi.org/10.1109/TVCG.2019.2934654

Topic Modeling algorithms are widely applied in digital humanities and computational social sciences to thematically segment corpora. However, as this task is highly subjective and domain-dependant, the results of these models usually require refinement and personalization. In this talk, I reflect on a journey of developing a series of visual analytics techniques that enable domain experts to externalize their domain knowledge and understanding for the task of topic model refinement. The talk highlights the open research questions, related to trusting machine learning models, that came up during the development process, namely:

- How much should we show the users?
- How much control should we give users?
- Do more confident decisions lead to better models?
- Does reporting that they trust the ML model means that the users understand its inner workings?
- Can we trust the users to interact directly with the ML models?
- Should we only care about the users trusting the ML model or also about the ML model trusting the users?
- How does explainability affect trust?

3.2 Trustworthy AI – A Medical Perspective

Jörn Kohlhammer (Fraunhofer IGD – Darmstadt, DE)

License \bigcirc Creative Commons BY 3.0 Unported license

© Jörn Kohlhammer

Joint work of Partly: Jörn Kohlhammer, Jan Burmeister, Jürgen Bernard, Thorsten May, Stefan Wesarg

Main reference Jan Burmeister, Jürgen Bernard, Thorsten May, Jörn Kohlhammer: "Self-Service Data Preprocessing and Cohort Analysis for Medical Researchers", in Proc. of the 2019 IEEE Workshop on Visual Analytics in Healthcare, VAHC 2019, Vancouver, BC, Canada, October 20, 2019, pp. 17–24, IEEE, 2019.

URL http://dx.doi.org/10.1109/VAHC47919.2019.8945040

This talk touched upon the experiences of the health sector with AI to foster the discussion on the requirements and expectation towards trust-building measures. The health sector has been the target of the AI technology field for many years, however resulting in more

40 20382 – Interactive Visualization for Fostering Trust in AI

frustration than benefit. There are some notable exceptions in radiology and other imagerelated areas. Overall though, there is a strong demand for trustworthy-AI approaches, be it organizational, psychological or technical flavors of such approaches. One observation is the granularity of AI techniques that should be small as part of healthcare processes. Another is the need for active involvement schemes to give medical experts some steering possibilities. Overall, there are many open questions, some of which can certainly be answered in this seminar.

3.3 Visual Analytics for Large-Scale ML Systems

Minsuk Khang (Oregon State University, Corvallis, US)

In this talk, I presented my research on creating visual analytics tools for interpreting machine learning (ML) systems that use very large datasets. I started by introducing ActiVis, a highly-cited paper and one of the early works in visualization for ML, which we designed and developed for engineers and data scientists at Facebook to interpret their industry-scale deep learning models. Then I presented several recent works to show how visual analytics tools can be integrated into an end-to-end ML workflow (e.g., model selection, fairness auditing). Lastly, I discussed my vision to further promote interactions between human and AI and broaden people's access to AI technologies, which can help build their trust in AI.

3.4 Autonomy and trust in spaceflight operations

Scott Davidoff (NASA JPL)

License O Creative Commons BY 3.0 Unported license O Scott Davidoff

This talk describes the results from a number of efforts to infuse explainable autonomous systems into spaceflight operations at the NASA Jet Propulsion Laboratory. The talk looks at how scientist- and engineer- users jobs change when autonomous spacecraft are introduced to mission operations contexts, and focuses on an example where visual analytics was used to summarize thousands of simulations, as a path to build trust by growing an understanding of how the spacecraft would respond to a wide variety of unexpected circumstances. The talk also uses Jonathan Grudin's framework for analyzing how CSCW applications fail, to to consider how autonomy transforms experts into novices, mis-aligning individual and organizational incentives, and following a dark pattern in the history of enterprise application failures.

3.5 ExplainExplore – Experimenting with Explanations

Jarke J. van Wijk (TU Eindhoven, NL)

License
 Creative Commons BY 3.0 Unported license
 I Jarke J. van Wijk

 Joint work of Dennis Collaris, Jarke van Wijk
 Main reference Explanations", in Proc. of the 2020 IEEE Pacific Visualization Symposium, PacificVis 2020, Tianjin, China Lean 2, 5, 2020.

After a short overview of work at TU Eindhoven on explainable AI, I presented the work of my PhD student Dennis Collaris. Using a surrogate model for explaining a single case is a popular approach in XAI, with LIME as the key example. However, this requires quite some parameter values to be chosen and also does not provide a global overview. Dennis has developed ExplainExplore: an interactive tool to explore explanations for complex models, enabling users to get an impression which features are important under what conditions. Also, a video, the paper, and a demo can be found at http://explaining.ml.

China, June 3-5, 2020, pp. 26-35, IEEE, 2020. URL https://doi.org/10.1109/PacificVis48177.2020.7090

Participants

 Michael Behrisch Utrecht University, NL
 Rita Borgo King's College London, GB

Mennatallah El-Assady
 Universität Konstanz, DE

Daniel A. Keim
Universität Konstanz, DE
Jörn Kohlhammer
Fraunhofer IGD –
Darmstadt, DE
Daniela Oelke
Hochschule Offenburg, DE

Maria Riveiro
 Jönköping University, SE

Tobias Schreck TU Graz, AT

Jarke J. van Wijk TU Eindhoven, NL

