

Interactive Visualization for Fostering Trust in ML

Polo Chau^{*1}, Alex Endert^{*2}, Daniel A. Keim^{*3}, and Daniela Oelke^{*4}

1 Georgia Institute of Technology – Atlanta, US. polo@gatech.edu

2 Georgia Institute of Technology – Atlanta, US. endert@gatech.edu

3 Universität Konstanz, DE,. keim@uni-konstanz.de

4 Hochschule Offenburg, DE. daniela.oelke@hs-offenburg.de

Abstract

The use of artificial intelligence continues to impact a broad variety of domains, application areas, and people. However, interpretability, understandability, responsibility, accountability, and fairness of the algorithms' results – all crucial for increasing humans' trust into the systems – are still largely missing. The purpose of this seminar is to understand how these components factor into the holistic view of trust. Further, this seminar seeks to identify design guidelines and best practices for how to build interactive visualization systems to calibrate trust.

Seminar August 28–September 2, 2022 – <http://www.dagstuhl.de/22351>

2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Computing methodologies → Machine learning; Human-centered computing → Visualization

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

Digital Object Identifier 10.4230/DagRep.12.8.103

1 Executive Summary

Polo Chau (Georgia Institute of Technology – Atlanta, US)

Alex Endert (Georgia Institute of Technology – Atlanta, US)

Daniel A. Keim (Universität Konstanz, DE)

Daniela Oelke (Hochschule Offenburg, DE)

License © Creative Commons BY 4.0 International license

© Polo Chau, Alex Endert, Daniel A. Keim, and Daniela Oelke

Artificial intelligence (AI), and in particular machine learning (ML) algorithms, are of increasing importance in many application areas. However, interpretability, understandability, responsibility, accountability, and fairness of the algorithms' results – all crucial for increasing humans' trust into the systems – are still largely missing. All major industrial players, including Google, Microsoft, and Apple, have become aware of this gap and recently published some form of Guidelines for the Use of AI. While it is clear that the level of trust in AI systems does not only depend on technical but many other factors, including sociological and psychological factors, interactive visualization is one of the technologies that has strong potential to increase trust into AI systems. In our Dagstuhl Seminar, we discussed the requirements for trustworthy AI systems including sociological and psychological aspects as well as the technological possibilities provided by interactive visualizations to increase human trust in AI. As a first step, we identified the factors influencing the organizational and sociological as well as psychological aspects of AI and partitioned them into relationship-based and evidence-based aspects. Next, we collected measures that may be used to approximate

* Editor / Organizer



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Interactive Visualization for Fostering Trust in ML, *Dagstuhl Reports*, Vol. 12, Issue 8, pp. 103–116

Editors: Polo Chau, Alex Endert, Daniel A. Keim, and Daniela Oelke



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

these aspects, such as interaction logs, eye tracking, and EEG. We also discussed the mechanisms to calibrate trust and their potential misuse. Finally, we considered the role that visualizations play in increasing trust in AI systems. This includes questions such as: Which mechanisms exist to make AI systems trustworthy? How can interactive visualizations contribute? Under which circumstances are interactive visualizations the decisive factor for enabling responsible AI? And what are the research challenges that still have to be solved – in the area of machine learning or interactive visualization – to leverage this potential in real world applications?

The seminar started with four keynote talks by experts in cognitive psychology, sociology, AI, and visualization, to provide participants with diverse perspectives that helped seed discussion topics. Then, the group decided to build 6 smaller groups to discuss the individual topics that should be worked on during the rest of the week. The six groups collectively came up with a longer list of potential topics surrounding the aspects of trust and machine learning. This list was voted on the plenum to distill it to the following four breakout groups: (1) Good practices and evil knobs in machine learning; (2) Evaluation, measures and metrics for trust in ML; (3) Interaction, expectations and dimension reduction; and (4) Definitions, taxonomy and relationships of trust in ML.

The outcome of this seminar is a better understanding of which aspects of trust have to be considered in fostering trust in AI systems and how interactive visualizations can help foster trust in artificial intelligence systems by making them more understandable and responsible. This will encourage innovative research and help to start joint research projects tackling the issue. Concrete outcomes are drafts of position papers describing the findings of the seminar and in particular, the research challenges identified in the seminar.

2 Table of Contents

Executive Summary

Polo Chau, Alex Endert, Daniel A. Keim, and Daniela Oelke 103

Overview of Talks

Visual, Interactive, and Explainable AI: Perspectives on Trust-Building through Explainability <i>Mennatallah El-Assady</i>	106
Cognitive perspectives on visualization and trust <i>Brian D. Fisher</i>	106
Visualizing Deep Networks <i>Barbara Hammer</i>	107
When Does Seeing Become Believing? Potential Impacts of Model Characteristics and Visual Cues on Human Decisions <i>Laura Matzen</i>	107
Relations between models, trust and processes involved in Machine Learning <i>Daniela Oelke</i>	108
Expectations, trust, and evaluation <i>Maria Riveiro</i>	109
Nonlinear dimensionality reduction – visualization with machine learning <i>Michel Verleysen</i>	109
A Computer Scientist’s Existential Crisis <i>Emily Wall</i>	110

Working groups

Definitions, taxonomy and relationships <i>Emma Beauxis-Aussalet, Peer-Timo Bremer, Steffen Koch, Jörn Kohlhammer, and Daniela Oelke</i>	110
A human-centered perspective on trust in AI-driven socio-technical systems <i>Peer-Timo Bremer, Emma Beauxis-Aussalet, Polo Chau, David S. Ebert, Daniel A. Keim, Steffen Koch, and Daniela Oelke</i>	112
Trust Junk and Evil Knobs: Duality of Trust-Calibration Design Measures <i>Alex Endert, Rita Borgo, Polo Chau, Mennatallah El-Assady, Laura Matzen, Adam Perer, Harald Schupp, Hendrik Strobelt, and Emily Wall</i>	113
Trust Evaluation <i>Maria Riveiro, Michael Behrisch, Simone Braun, David S. Ebert, Daniel A. Keim, Tobias Schreck, and Hendrik Strobelt</i>	114
The Flow of Trust: An Interactive Visualization Framework for Externalizing, Exploring and Explaining Trust in ML Applications <i>Stef Van den Elzen, Gennady Andrienko, Natalia V. Andrienko, Brian D. Fisher, Rafael M. Martins, Jaakko Peltonen, Alexandru C. Telea, and Michel Verleysen</i>	115

Participants 116

3 Overview of Talks

3.1 Visual, Interactive, and Explainable AI: Perspectives on Trust-Building through Explainability

Mennatallah El-Assady (ETH Zürich, CH)

License © Creative Commons BY 4.0 International license
 © Mennatallah El-Assady
URL <https://el-assady.com/>

Interactive, mixed-initiative machine learning promises to combine the efficiency of automation with the effectiveness of humans for collaborative decision-making and problem-solving process. This can be facilitated through co-adaptive visual interfaces.

In the first part of this talk, I recapped the definitions of mixed-initiative analysis, arguing for the need for effective explanations, both from the side of the human as well as the AI agents. Next, I summarized visual, interactive, and explainable AI approaches along the process of understanding, diagnosis, and refinement of models.

Second, I reviewed human-centered evaluations in human-centered AI, focusing on how the trustworthiness of AI models and the trustworthiness of explanations were evaluated in previous works. Following up on this, I presented the output of the survey on enhancing trust in machine learning through visualization.

Lastly, I ended the talk with some reflections and open questions relating to trust-building through explainability.

3.2 Cognitive perspectives on visualization and trust

Brian D. Fisher (Simon Fraser University – Surrey, CA)

License © Creative Commons BY 4.0 International license
 © Brian D. Fisher
Joint work of Brian D. Fisher, Samar Al-Hajj, Richard Arias-Hernández, Linda T. Kaastra
Main reference B. Fisher & D. Kasik (2023) Pair Analytics in a Visual Analytics Context. Proceedings of the 56th Annual Hawaii International Conference on System Sciences. IEEE Digital Library, to appear.

Highly Integrated Basic and Application- Responsive (HIBAR) research approaches try to find optimal ways to bridge the knowledge-creation of basic science with the design and engineering of advanced technologies. Here I discuss ways to apply this approach to build interactive visualization systems that might be used to establish trust in machine learning processes that are grounded in basic research in the cognitive science of visually-enabled reasoning and agent-agent communication and at the same time to contribute to knowledge about those processes. I begin with a discussion of various kinds of trust and ways in which they are cognitively processed. In order to reduce the complexity of this analysis I use David Marr's triune approach from Vision (1984) of implementation, algorithm, and operational requirements and its extension by Poggio to include perspectives on the evolution and development of expertise in a given cognitive task.

This analysis is helpful in design of visualization for many complex cognitive processes that are supported by visual information in structured environments, such as our new project on cancer diagnosis through medical image analysis using machine learning. In order to address issues of trust, we must build a parallel understanding of how people are able to coordinate their behaviour with that of other agents. I build this from H.H. Clark's psycholinguistic pragmatic approach to human-human coordination in Joint Activities. Here

too we see benefits from Marr’s Triune approach, with D’Andrade’s Cognitive Anthropology and Hutchins’ approach to Cognitive Ethnography as examples of ways in which groups of people, their technologies, and channels of communication interact to produce extended and distributed cognitive systems, with examples from work in my laboratory using our Pair Analytics and Group Analytics approach to safety and health decision-making, and our recent collaboration on Decision Intelligence approached pioneered by Lorien Pratt.

My final topic is pragmatic—how can we most effectively build HIBAR research programs that bridge real-world applications and creation of scientific knowledge. I briefly discuss new approaches to technoscience and creative design collaboration that require a rethinking of the structures that define research organizations in the university and industry. In order to build these systems in a reflective manner we must take the lens we used to understand other organizations as distributed systems and apply it to our own organizations.

3.3 Visualizing Deep Networks

Barbara Hammer (Universität Bielefeld, DE)

License © Creative Commons BY 4.0 International license
© Barbara Hammer

Joint work of Alexander Schulz, Fabian Hinder, Barbara Hammer

Main reference Alexander Schulz, Fabian Hinder, Barbara Hammer: “DeepView: Visualizing Classification Boundaries of Deep Neural Networks as Scatter Plots Using Discriminative Dimensionality Reduction”, in Proc. of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, pp. 2305–2311, ijcai.org, 2020.

URL <http://dx.doi.org/10.24963/ijcai.2020/319>

In this spotlight talk, a pipeline for visualizing the classification boundary of deep networks together with example data has been presented. The technology relies on two tricks: using discriminative dimensionality reduction to shape the otherwise ill-posed problem of dimensionality reduction from high dimensional spaces according to the task at hand, and sampling in the projection instead of the data space for efficiency. For deep networks, the first part can be approximated using one backpropagation loop only. The algorithm can be substantiated by convergence guarantees, and it is available as code <https://github.com/LucaHermes/DeepView>

3.4 When Does Seeing Become Believing? Potential Impacts of Model Characteristics and Visual Cues on Human Decisions

Laura Matzen (Sandia National Labs – Albuquerque, US)

License © Creative Commons BY 4.0 International license
© Laura Matzen

Joint work of Laura Matzen, Breannan C. Howell, Zoe Gastelum, Kristin M. Divis, Michael C. Trumbo

How can interactive visualizations foster trust in machine learning (ML)? Trust is an extremely complex issue. A user may fail to trust a model when they could benefit from doing so, or they might trust too much, complying with a model’s outputs even when they are erroneous. How can we support people in developing appropriate levels of trust in an ML tool, so that the human-machine system can reach the highest possible level of performance? In this talk, I discuss some of the cognitive issues that come into play when humans are asked to make decisions based on visualizations and other representations of information. I present two

lines of research, one focused on the impact of ML errors on human decision making and one focused on visualizations of state uncertainty. Across these two sets of experiments, we found that different representations of the same information can lead to different patterns of decisions. People’s ability to detect ML errors is impacted by the overall error rate of the system. Their tolerance of risk in a decision making task is impacted by the way in which information about risk is presented. In addition, individual differences in cognition and domain expertise influence participants’ interpretation of and trust in ML outputs. The results of these experiments illustrate some of the many factors that can influence users’ trust and decisions. Additional research at the intersection of cognitive science, data science, data visualization, and visual analytics will be required to develop a systematic understanding of these factors and the interactions between them.

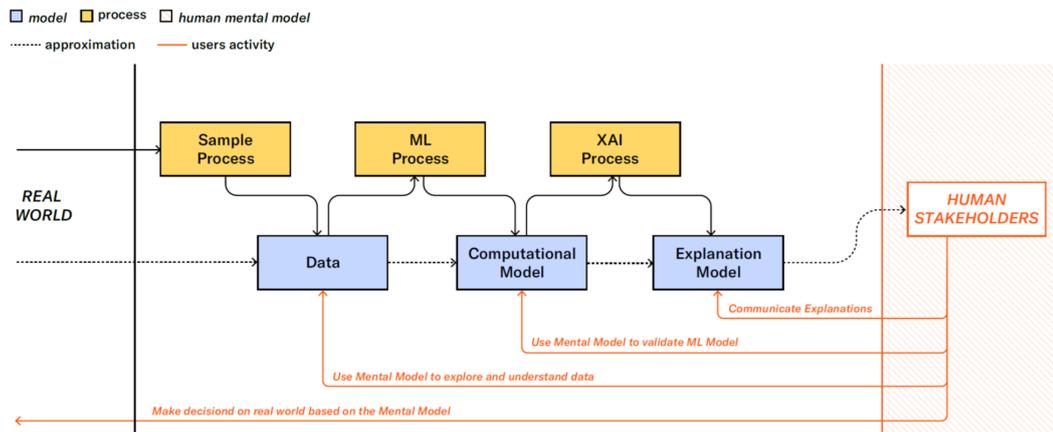
3.5 Relations between models, trust and processes involved in Machine Learning

Daniela Oelke (Hochschule Offenburg, DE)

License  Creative Commons BY 4.0 International license
© Daniela Oelke

Joint work of the participants of the Dagstuhl Seminar 19452 “Machine Learning Meets Visualization to Make Artificial Intelligence Interpretable”

About three year ago a related seminar entitled “Machine Learning Meets Visualization to Make Artificial Intelligence Interpretable” took place in Dagstuhl. This talk introduces one of the outputs of this Dagstuhl Seminar, a diagram illustrating the relations between the different types of model and processes involved in the ML process and its explanation including the interactions of the human stakeholders with the models. The diagram does not yet address issues with trust, but could be adapted towards this direction.



3.6 Expectations, trust, and evaluation

Maria Riveiro (Jönköping University, SE)

License © Creative Commons BY 4.0 International license
© Maria Riveiro

Main reference Maria Riveiro, Serge Thill: ““That’s (not) the output I expected!” On the role of end user expectations in creating explanations of AI systems”, *Artif. Intell.*, Vol. 298, p. 103507, 2021.

URL <http://dx.doi.org/10.1016/j.artint.2021.103507>

This talk focuses on the role of expectations in designing explanations from Artificial Intelligence/Machine Learning (AI/ML) -based systems. Explanations are crucial for system understanding that, in turn, are very relevant to supporting trust and trust calibration in such systems. I discuss the connections between expectations, explanations and trust in human-AI/ML system interaction.

I present two recent studies ([1, 2]) investigating if expectations modulate what people want to see and when from an AI/ML system when carrying out analytical tasks.

We found out that,

- For matched expectations, an explanation is often not required at all, while if one is, it is of the factual type
- For mismatched expectations, the picture is less clear, primarily because there does not seem to be a unique strategy, although mechanistic explanations are requested more often than other types

Overall, user expectations are a significant variable in determining the most suitable content of explanations (including whether an explanation is needed at all). More research is needed to investigate the relationship between expectations and explanations, and how they support trust calibration.

References

- 1 Riveiro, M., and Thill, S. (2021). That’s (not) the output I expected! On the role of end user expectations in creating explanations of AI systems. *Artificial Intelligence*, 298, 103507.
- 2 Riveiro, M., and Thill, S. (2022). The challenges of providing explanations of AI systems when they do not behave like users expect. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 110-120).

3.7 Nonlinear dimensionality reduction – visualization with machine learning

Michel Verleysen (University of Louvain, BE)

License © Creative Commons BY 4.0 International license
© Michel Verleysen

Joint work of Michel Verleysen, John A. Lee, Cyril de Bodt, Pierre Lambert

Nonlinear dimensionality reduction (NLDR) is a branch of the wide machine learning field. NLDR is essentially unsupervised, which means that it is used to find something (information) in data, but we do not know in advance what kind of information. Consequently, even if it is easy to agree on the general principle of reducing the dimension of data without losing too much information, it is very difficult to agree on a scientific measure of how this loss is evaluated, leading to hundreds of NLDR methods. None of them can be objectively considered as better than others; they all reflect a specific user’s point of view. Popular methods are often those that come with an efficient implementation, rather than being chosen for the quality (seen by the user) of their outputs.

This talk insists on one of the users’ point of views, often underestimated in the literature: the compromise between a global and a local mapping of the data. We show that multiscale methods may produce much more interesting representations, with an additional computational cost that can be limited with the development of fast yet accurate algorithms.

3.8 A Computer Scientist’s Existential Crisis

Emily Wall (Emory University – Atlanta, US)

License  Creative Commons BY 4.0 International license
© Emily Wall

This talk began with the observation that trust in AI systems encompasses trust in (1) the model – that it is accurate and just, (2) the decision maker – that the human-in-the-loop standing between you and the model decision or prediction has your best interest in mind, and (3) oneself – that you are equipped to make informed decisions. The technical aspects of trust in AI systems are actually only a small part of the story. Improving model accuracy and quantifying and mitigating bias in models can serve to calibrate trust. For visualization researchers in particular, this begs the question: what is our role? Where can we have impact? The talk asserts that this is a challenging socio-technical problem, requiring a suite of methodologies and frameworks that are not especially common in our community. I conclude the talk with a reference to a paper[1] that leverages a valuable socio-technical perspective to coin the concept “human-centered explainable AI” abbreviated HCXAI. This paper introduces important frameworks (including critical technical practice, reflective design, value-sensitive design) that can serve as a starting point for visualization researchers to expand their tool belts to include critical socio-technical frameworks to inform next steps addressing trust in AI through interactive visualization.

References

- 1 Upol Ehsan and Mark O. Reidl (2020). *Human-centered explainable ai: Towards a reflective sociotechnical approach*. International Conference on Human-Computer Interaction.

4 Working groups

4.1 Definitions, taxonomy and relationships

Emma Beauxis-Aussalet (VU University Amsterdam, NL), Peer-Timo Bremer (LLNL – Livermore, US), Steffen Koch (Universität Stuttgart, DE), Jörn Kohlhammer (Fraunhofer IGD – Darmstadt, DE), and Daniela Oelke (Hochschule Offenburg, DE)

License  Creative Commons BY 4.0 International license
© Emma Beauxis-Aussalet, Peer-Timo Bremer, Steffen Koch, Jörn Kohlhammer, and Daniela Oelke

A major point of discussion was the definition of trust. It quickly became clear that we needed to differentiate between (1) trust (as historically defined by philosophy) that denotes the relationship between humans and between humans and organizations, and (2) trust that concerns technical artifacts, especially machine learning components. A first attempt was the separation of trust on the human side and confidence on the technical side. While this worked to structure the terms that are related to trust, the current (different) use of these terms in the

various communities led to reconsiderations. A strong second candidate was the separation into subjective trust and objective trust. However, this did not transport well, that trust is inherently between humans, not a state of mind of a single human. Also, the objectivity of several aspects on the technical side, when it comes to trusting a technical artifact, was not adequately covered. An extensive research of terms by Emma Beauxis-Aussalet led us to the final two terms that we now use in this Dagstuhl seminar: relationship-based trust on the human side, and evidence-based trust on the technical side. The material that we prepared also contains a diagram showing how these two sides relate to each other during human decision making.

Detailed discussion of the definition of the term “trust”

Prior art [2, 1, 3] suggests that reliance on relationships is the core framework of trust, whether it concerns trust in another human, in an institution, in oneself, or in technology – the latter three being modeled after the first. This reliance can be warranted, based on evidence, but also arises from the necessity to depend on another party, which is potentially fallible. Evidence may eliminate the risks of such dependency (e.g., *well-grounded trust*) or only reduce them and demonstrate the value of such dependency (e.g., *justified trust*).

A body of evidence can inform the decision to enter or exit a relationship of dependency with another party, human or object. But such evidence does not only concern technical information, e.g., about the reliability of a ML system. It also concerns organisational and societal considerations that are external to the trustworthiness of technology. It is thus essential to consider the relationships that arise from integrating technology in human organisations and societies.

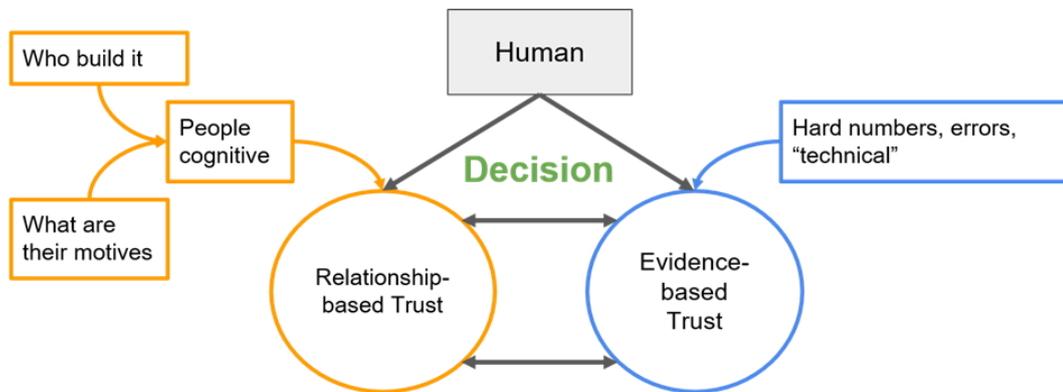
In essence, **trust may not be achieved solely by providing evidence on technical risks**: Trust can be established or withdrawn for reasons other than the quality of a technology, for instance due to material necessities or power dynamics. Trust in technology is also relationship-based because i) it relies on the relationships between a larger body of actors who have agency on the technology or its usage, ii) humans also establish relationships of reliance to the technology. The latter is arguably about trust, as the motives of the trustee is an essential component of trust [2] and technology itself has no motives. However, we can acknowledge that anthropomorphism occurs, and includes the illusory attribution of motives to AI and automated systems. Hence, computer science and visualization research aiming to support trust in machine learning must consider the cognitive, emotional, and societal aspects that are inherent to relationship-based trust.

Recent works also discuss trust by persons or organisations in AI and Machine Learning [4, 5]. While these works describe facets that can be used to make a distinction regarding relation-based and evidence based trust, this is not discussed explicitly.

Evidence-based trust is built on factual information such as error metrics, uncertainty estimates, and all the information available to reach a decision or complete a task. This information is measurable and verifiable. Yet humans are not entirely objective and base their decisions on a combination of subjective and objective aspects of a situation: trust in humans, organizations, or technology may not be fully informed by evidence, but combine relationship-based and evidence-based trust. Furthermore, a single human is not able to fully assess all possible evidence. Thus relationship-based trust is necessary to mediate the complexity of technology by delegating their assessment and management to a network of specialised actors.

Relationship-based trust is built on practical cooperations between actors with specific skills, motives, and will. In machine learning, the parties and roles of relationship-

based trust are largely evolving, same as the role of webmaster at the beginning of the web eventually evolved to a network of specialists. However evolutive the relationships, to support relationship-based trust it is essential to identify the goals, tasks, information needs, and profile of each actor.



References

- 1 Aneil K. Mishra. “Organizational Responses to Crisis: The Centrality of Trust.” In: *Trust in Organizations*. Ed. by Roderick M. Kramer and Thomas Tyler. Newbury Park, California, USA: Sage, 1996
- 2 Carolyn McLeod, “Trust”, *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), Edward N. Zalta (ed.), <https://plato.stanford.edu/archives/fall2021/entries/trust/>
- 3 Rotter, Julian B. “A new scale for the measurement of interpersonal trust” *Journal of personality* 35.4 (1967): 651-665.
- 4 Toreini, Ehsan, et al. “The relationship between trust in AI and trustworthy machine learning technologies.” *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020.
- 5 Ashoori, Maryam, and Justin D. Weisz. “In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes.” *arXiv preprint arXiv:1912.02675* (2019).

4.2 A human-centered perspective on trust in AI-driven socio-technical systems

Peer-Timo Bremer (LLNL – Livermore, US), Emma Beauxis-Aussalet (VU University Amsterdam, NL), Polo Chau (Georgia Institute of Technology – Atlanta, US), David S. Ebert (University of Oklahoma – Norman, US), Daniel A. Keim (Universität Konstanz, DE), Steffen Koch (Universität Stuttgart, DE), and Daniela Oelke (Hochschule Offenburg, DE)

License © Creative Commons BY 4.0 International license

© Peer-Timo Bremer, Emma Beauxis-Aussalet, Polo Chau, David S. Ebert, Daniel A. Keim, Steffen Koch, and Daniela Oelke

Trust in the information provided is often cited as one of the key challenges to fully integrate AI-driven systems into high consequence decisions. However, there rarely exist a clear definition of the concept, what can be done to influence trust, and even whether the implicit goal to increase trust is always appropriate. The HCI community is often asked to mediate between the various stakeholders and visualization in particular is seen as the ideal conduit

to convey both subjective and objective information. However, the amalgamation of human and social aspects with more technical concerns on correctly conveying unbiased information makes formulating a clear research perspective difficult. Here we argue that the general notion of “trust” should be thought of as two related but fundamentally different concepts: relationship-based trust and evidence-based trust. A typical example of the former is the trust one places in an car functioning, which is rooted in the believe that an engineer certified the system rather than in any personal knowledge of the mechanics. Conversely, evidence-based trust is based in factual information, i.e., statistics on past performance or uncertainty bounds, analyzed directly. In both cases, the overarching goal should be to correctly calibrate trust to avoid both unfounded over-trust, for example, based on the social network echo chamber, as well as unjustified skepticism. This perspective will first define both concepts, show how they implicitly or explicitly align with prior arguments, and that they lead to fundamentally different research challenges. Subsequently, the perspective discusses priority research directions aimed at calibrate both form of trust in AI-driven system, how the different notions interact, and most importantly areas where unfettered research may raise ethical concerns. While this perspective grew out of discussions in the HCI community addressing many of the challenges will require convergent research in cognitive science, machine learning, ethics, visualization, and many others.

4.3 Trust Junk and Evil Knobs: Duality of Trust-Calibration Design Measures

Alex Endert (Georgia Institute of Technology – Atlanta, US), Rita Borgo (King’s College London, GB), Polo Chau (Georgia Institute of Technology – Atlanta, US), Mennatallah El-Assady (ETH Zürich, CH), Laura Matzen (Sandia National Labs – Albuquerque, US), Adam Perer (Carnegie Mellon University – Pittsburgh, US), Harald Schupp (Universität Konstanz, DE), Hendrik Strobelt (MIT-IBM Watson AI Lab – Cambridge, US), and Emily Wall (Emory University – Atlanta, US)

License © Creative Commons BY 4.0 International license
© Alex Endert, Rita Borgo, Polo Chau, Mennatallah El-Assady, Laura Matzen, Adam Perer, Harald Schupp, Hendrik Strobelt, and Emily Wall

Many AI systems make claims that specific design choices enhance trust or serve to calibrate trust. However, interface design choices are not neutral with respect to trust. There is inherent duality – that the same design choice may enhance trust in some cases, while simultaneously detracting in others. This group conceptualized “trust junk,” analogous to “chart junk” in visualization, i.e., design choices intended to enhance trust without any specific connection to data, model, or the nature of the decision.

Consider AIs that utilize social information, e.g., “5 others in the organization have accepted a recommendation today.” This choice may increase trust in the AI having social endorsement by others; however, this may also be used to create unfair social pressure and manipulate choices of the user that serve the interface creators. The group expanded these examples to consider different kinds of “knobs,” which represent different design choices that can be made in the creation of an AI system. These include choices about which datasets are modeled, how the outputs of the system are represented to users, and what options users have for interacting with the outputs, the model, or the data. Turned in one direction in a given context, these knobs may enhance trust in the system. When considered from an adversarial perspective, these choices could also be used to mislead users or to promote

unwarranted levels of trust in a system. We construct a framework of these knobs and assert that understanding this space necessitates a sociotechnical approach; concrete generalizable interface guidelines may not yet be made.

4.4 Trust Evaluation

Maria Riveiro (Jönköping University, SE), Michael Behrisch (Utrecht University, NL), Simone Braun (Hochschule Offenburg, DE), David S. Ebert (University of Oklahoma – Norman, US), Daniel A. Keim (Universität Konstanz, DE), Tobias Schreck (TU Graz, AT), and Hendrik Strobelt (MIT-IBM Watson AI Lab – Cambridge, US)

License  Creative Commons BY 4.0 International license

© Maria Riveiro, Michael Behrisch, Simone Braun, David S. Ebert, Daniel A. Keim, Tobias Schreck, and Hendrik Strobelt

Trust assessment during the data analysis process is a challenging task. Only if stakeholders have trust in the used data, algorithmic, and visual-interactive components, the results will be accepted, relied on and applied. In an ideal system, the trust of the user could be observed (measured), and the system could be adapted to increase trust where it is lacking, e.g., by providing additional information, explanations, or summarizations. However, trust is dynamic and emerges due to many influencing factors, both from the data analysis system designs, as well as personal factors. Additionally, learning effects, change in the user tasks, or socio-cultural influences complicate trust calibration. To date, several trust scales exist, but it is hard to assess how technological aspects of ML/AI systems affect trust and how to carry out empirical evaluations that isolate the effects that changes in technology have on trust.

Related work in Human-centered AI and Human-computer interaction suggests frameworks for compartmentalizing trust into cognition-based and affect-based trust. We combine two earlier frameworks from Madsen & Gregor (2000) [1] and Gulati et al. (2019) [2] and add socio-cultural trust factors that were not considered in the discussion. We scrutinize these frameworks considering particularities of AI/ML, Visualization and Interaction. Overall, we propose the following aspects/constructs of trust:

- Cognition-Based Trust
 - Perceived risk
 - Benevolence
 - Competence
 - Reciprocity
- Affect-Based Trust
 - Faith/Confidence
 - Personal Attachment
 - Personality (locus of control, risk-averseness, etc.)
 - Experience (past experiences)
- Social and cultural trust (environmental and contextual factors)
 - Peer Influence
 - Crowd Behavior
 - Cultural Norms

We complement our work by discussing which possible proxies, indirect measures that allow us to approximate trust scores, we can use to assess these aspects, and elaborate on which proxies are most suitable for each trust aspect/construct. The measures outlined are

questionnaires, EEG, Eye-Tracking, Face expressions, movement, recognition (e.g. FACS), Galvanic skin response (e.g. glove), Interaction logs, Feedback from users, Voice recognition, changes (e.g. shimmer, jitter, pitch, balance via GeMAPS), Model questioning users, Games and A/B-Testing. Future work has to prove their applicability for this challenging task.

We finalize with a discussion on how scalable and how good proxies these measures are for trust, and we outline the next steps in utilizing this knowledge for carrying out empirical evaluations and during the design process of Vis/ML/AI-based systems.

References

- 1 Madsen, M. and Gregor, S., 2000. Measuring human-computer trust. In 11th Australasian conference on information systems (Vol. 53, pp. 6-8). Brisbane, Australia: Australasian Association for Information Systems.
- 2 Gulati, S., Sousa, S. and Lamas, D., 2019. Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, 38(10), pp.1004-1015.

4.5 The Flow of Trust: An Interactive Visualization Framework for Externalizing, Exploring and Explaining Trust in ML Applications

Stef Van den Elzen (TU Eindhoven, NL), Gennady Andrienko (Fraunhofer IAIS – Sankt Augustin, DE), Natalia V. Andrienko (Fraunhofer IAIS – Sankt Augustin, DE), Brian D. Fisher (Simon Fraser University – Surrey, CA), Rafael M. Martins (Linnaeus University – Växjö, SE), Jaakko Peltonen (Tampere University of Technology, FI), Alexandru C. Telea (Utrecht University, NL), and Michel Verleysen (University of Louvain, BE)

License © Creative Commons BY 4.0 International license

© Stef Van den Elzen, Gennady Andrienko, Natalia V. Andrienko, Brian D. Fisher, Rafael M. Martins, Jaakko Peltonen, Alexandru C. Telea, and Michel Verleysen

Currently, trust in Machine Learning applications is an implicit process that takes place in the mind of the user. As a result there is no method of feedback or communication of trust that can be acted upon. Trust differs from mere ability to inspect the model and from a model's claimed confidence in its predictions. frameworks that support such aspects are not sufficient to support trust. We argue that trust needs to be considered as a first-class citizen in the workflow of developing and using machine learning models. We present a formalization of trust flow and externalization as part of interactive machine learning workflows for analysis and for decision-making. The formalization differentiates several user roles in exploring machine learning models at different workflow stages and their corresponding opportunities for explicit communication of trust targeted at these stages. The formalization enables construction of interaction modes and interfaces to help users to efficiently build and communicate trust in ways that are appropriate for a given stage in the analytic process. Moreover, the formalization differentiates the roles of model exploration and trust communication of the user as well as differentiating user trust from a model's internal probabilistic representations. We formulate several research questions and directions arising from our framework which include

- (a) typology/taxonomy of trust objects, trust issues, and possible reasons for (mis)trust;
- (b) formalisms to represent trust in machine-readable form;
- (c) ways for users to express their state of trust;
- (d) ways to explore and develop trust over models and their different aspects using visual interactive techniques;
- (e) ways to facilitate the user's expression and communication of the state of trust using visual interactive techniques.

Participants

- Gennady Andrienko
Fraunhofer IAIS –
Sankt Augustin, DE
- Natalia V. Andrienko
Fraunhofer IAIS –
Sankt Augustin, DE
- Emma Beauxis-Aussalet
VU University Amsterdam, NL
- Michael Behrisch
Utrecht University, NL
- Rita Borgo
King’s College London, GB
- Simone Braun
Hochschule Offenburg, DE
- Peer-Timo Bremer
LLNL – Livermore, US
- Polo Chau
Georgia Institute of Technology –
Atlanta, US
- David S. Ebert
University of Oklahoma –
Norman, US
- Mennatallah El-Assady
ETH Zürich, CH
- Alex Endert
Georgia Institute of Technology –
Atlanta, US
- Brian D. Fisher
Simon Fraser University –
Surrey, CA
- Barbara Hammer
Universität Bielefeld, DE
- Daniel A. Keim
Universität Konstanz, DE
- Steffen Koch
Universität Stuttgart, DE
- Jörn Kohlhammer
Fraunhofer IGD –
Darmstadt, DE
- Rafael M. Martins
Linnaeus University – Växjö, SE
- Laura Matzen
Sandia National Labs –
Albuquerque, US
- Daniela Oelke
Hochschule Offenburg, DE
- Jaakko Peltonen
Tampere University of
Technology, FI
- Adam Perer
Carnegie Mellon University –
Pittsburgh, US
- Maria Riveiro
Jönköping University, SE
- Tobias Schreck
TU Graz, AT
- Harald Schupp
Universität Konstanz, DE
- Hendrik Strobelt
MIT-IBM Watson AI Lab –
Cambridge, US
- Alexandru C. Telea
Utrecht University, NL
- Stef Van den Elzen
TU Eindhoven, NL
- Michel Verleysen
University of Louvain, BE
- Emily Wall
Emory University – Atlanta, US

