EDITOR: Theresa-Marie Rhyne, theresamarierhyne@gmail.com

DEPARTMENT: VISUALIZATION VIEWPOINTS

The Role of Interactive Visualization in Fostering Trust in AI

Emma Beauxis-Aussalet, *Vrije Universiteit Amsterdam, 1081 HV, Amsterdam, The Netherlands*
Michael Behrisch, *Utrecht University, 3512 BS, Utrecht, The Netherlands*
Rita Borgo, *King's College London, WC2B 4BG, London, United Kingdom*
Duen Horng Chau, *Georgia Tech, Atlanta, 30308, GA, USA*
Christopher Collins, *Ontario Tech University, Ontario, L1G 0C5, Canada*
David Ebert, *University of Oklahoma, Norman, OK, 73019, USA*
Mennatallah El-Assady, *University of Konstanz, 78464, Konstanz, Germany*
Alex Endert, *Georgia Tech, Atlanta, GA, 30332, USA*
Daniel A. Keim, *University of Konstanz, 78464, Konstanz, Germany*
Jörn Kohlhammer, *Fraunhofer IGD, 64283, Darmstadt, Germany*
Daniela Oelke, *Offenburg University, 77654, Offenburg, Germany*
Jaakko Peltonen, *Tampere University, FI-33014, Tampere, Finland*
Maria Riveiro, *Jönköping University, SE-551 11, Jönköping, Sweden*
Tobias Schreck, *Graz University of Technology, 8010, Graz, Austria*
Hendrik Strobelt, *IBM Research, Cambridge,MA, 02142, USA*
Jarke J. van Wijk, *Eindhoven University of Technology, 5612 AE, Eindhoven, The Netherlands*

The increasing use of artificial intelligence (AI) technologies across application domains has prompted our society to pay closer attention to AI's trustworthiness, fairness, interpretability, and accountability. In order to foster trust in AI, it is important to consider the potential of interactive visualization, and how such visualizations help build trust in AI systems. This manifesto discusses the relevance of interactive visualizations and makes the following four claims: i) trust is not a technical problem, ii) trust is dynamic, iii) visualization cannot address all aspects of trust, and iv) visualization is crucial for human agency in AI.

With the increased use of AI techniques, there are increased concerns about trustworthiness, fairness, interpretability, and accountability of these systems. Discussing and addressing these concerns is an inherently multi-disciplinary problem, requiring conversations and research activities that cross disciplines and methods. In order to foster trust in AI, it is important to consider the potential of interactive visualization, and how such visualizations help build trust in AI systems. This manifesto discusses the relevance of interactive visualizations in fostering trust in AI and makes the following four interrelated claims:

1) **Trust is not a technical problem.** Therefore, it is risky to address it without considering the

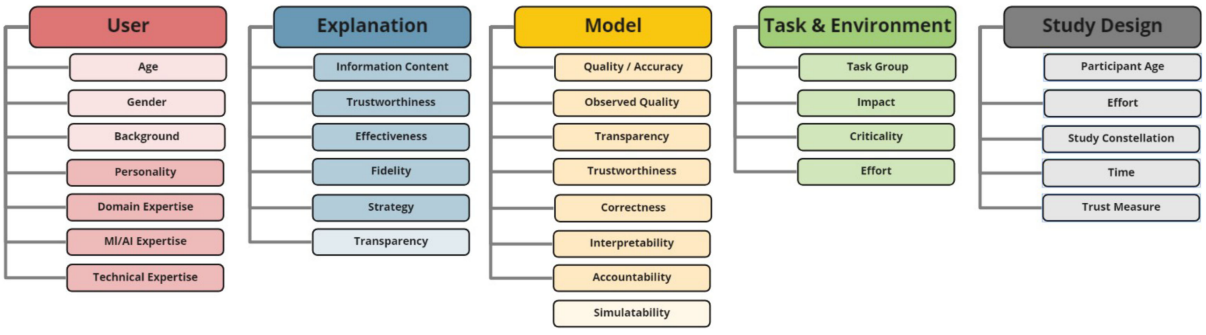

FIGURE 1. Design dimensions for AI systems include a wide spectrum of personal and environmental factors, derived from a structured literature survey.[1]

organizational, sociological, and psychological factors that affect trust.

2) **Trust is dynamic.** It needs to be evaluated and reconsidered regularly, as the world, and people's perception of it, evolves over time, and so does our trust toward AI.
3) **Visualization cannot address all aspects of trust.** Despite its strong potential in providing algorithmic and data transparency, visualization designs could be subjective and opinionated, and may not directly facilitate all factors that can help build trust—poor designs could even induce mistrust.
4) **Visualization is crucial for human agency in AI.** It is a powerful interface that bridges humans, data, and algorithms, and a flexible tool that can adapt to different users' evolving values, knowledge, and preferences.

Together, our four claims underscore that any successful methods to help people gain trust in AI must carefully consider and adapt to people's evolving needs and nontechnical factors that affect trust; while not a panacea, visualization is a powerful interface that empowers and bridges humans with AI to enable such essential consideration and adaptation. This manifesto provides a critical dissection of the important and nuanced problem of fostering trust in AI, and offers fresh interdisciplinary perspectives for inspiring future research.

Trust is Not a Technical Problem

Important sociological and psychological factors that affect trust are often neglected or not taken into account when designing and building AI-based visualization systems.[2] Those factors include, for instance, personal fears and values, unfairness, external actors (social and community), knowledge and experience, and others (Figure 1). As these challenges are inherently nontechnical, limiting their solution to technical design decisions does not fully resolve the issues. For example, some personal fears arise when there are misconceptions about what AI is or what it can achieve.[3] Similarly, end-users might have had negative experiences with AI in the past, or might be wary of the unclear accountability, responsibility, or level of human control. Hence, prior experiences with AI technologies, or a lack of such experiences, may shape one's perception of trustworthiness. Such prior experiences can lead to skepticism and doubts toward AI-based systems that need to be addressed on a nontechnical level, such as through targeted communication campaigns, affective design, or teaching. More research is needed to address such concerns; neglecting them can hinder trusting an otherwise trustworthy AI-based system.

Trust is Dynamic

Gaining trust in AI is a dynamic process. For high-stake applications, trust is usually developing slowly but can be lost quickly. Even if trust is once established, we cannot be sure that our trust would never be misplaced. Hence, trust has to be evaluated and reconsidered regularly. The need to re-evaluate trust can arise for technical reasons, such as shifts in data and error distributions, or changes in tasks' scope, goals, and practical impacts. Such technical issues can remain undiscovered at the time of deployment, and emerge over time or through user behaviors.

Trust is also a personal issue. Our human experience of trust can be subjective, e.g., depending on one's personal history, background, AI expertise, personality, cultural background, and social environment.[4] These factors also evolve over time and are influenced by our daily experiences, news, media, friends, family, and society. In addition, humans are prone to overtrust and

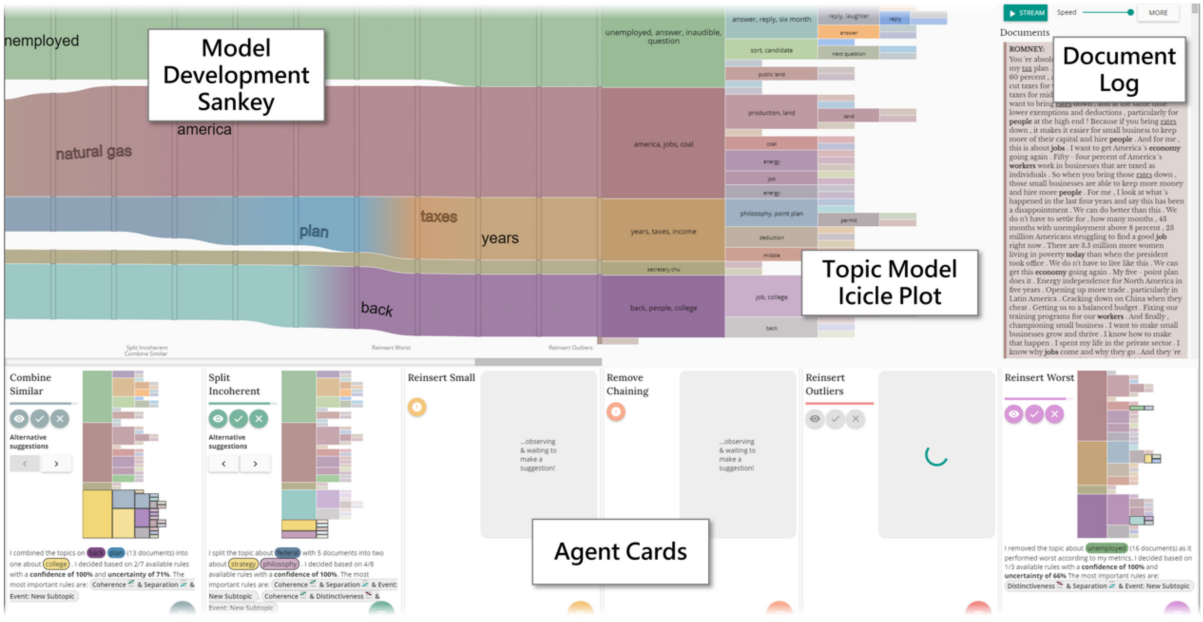

FIGURE 2. The topic modeling system by Sperrle *et al.*[11] supports users with single-objective guidance agents that, over time, learn in which analysis contexts to provide their specific suggestions of alternative models. This adaptation over time personalizes the system to a given user and their needs and expectations, thereby increasing their trust (see Claim 2).

distrust.[5] This means that any methods that aim at building trust need to be adaptive, to adjust to users' changing needs and trust levels, and to help users calibrate their trust to an appropriate level. For example, Sperrle *et al.*[11] developed a system for the guided refinement of topic models (Figure 2). As the system incrementally adds more documents, different single-objective guidance agents observe the process and make specific suggestions of alternative models. Over time, all agents learn in which analysis contexts their suggestions are typically accepted or rejected, and adapt their suggestions of alternative models to these contexts. The evaluation shows that several participants were initially skeptical toward the agents, but were willing to spend significant time training them (by providing relevance feedback), expecting better suggestions in the future. Similarly, the perception that residents have of tornado predictions and warnings can degrade over time when false positives occur (even when uncertainty or low likelihood is communicated).[12] Future research aiming to build trust in AI systems should help people decide if, when, and to what extent they should trust a specific system.

Visualization Cannot Address All Aspects of Trust

As stated above, trustworthy AI requires far more than technical solutions. The Guidelines for Trustworthy AI of the European Union,[6] as one influential set of guidelines among similar guidelines published by companies and public institutions alike, mention a broad range of challenges, including: human agency and oversight, in which legal frameworks are central; technical robustness and safety, for which technical solutions are requisite; and privacy and data governance, for which careful consideration and negotiation of individual and public interests is essential. Visualization systems that first and foremost take a technical approach, such as providing algorithmic transparency or revealing evidence from large datasets, cannot address such issues comprehensively.

When a decision is made, whether to trust a tool or not, social factors can play an important role as well. For instance, someone might trust an AI system because a friend or a trusted expert recommended it. In this case, the trust we have in a person has implications on our trust in the AI system. However, convincing someone to put trust in an AI system can also be the result of a communication process. Expressive visualizations can support that process by providing the necessary transparency or explaining the inner workings of a system. In this case, visualizations are a valuable tool but no standalone solution.[7]

Visualization can be opinionated. This can be problematic if it arises from benign design opinions to purposefully disguising information through agenda-

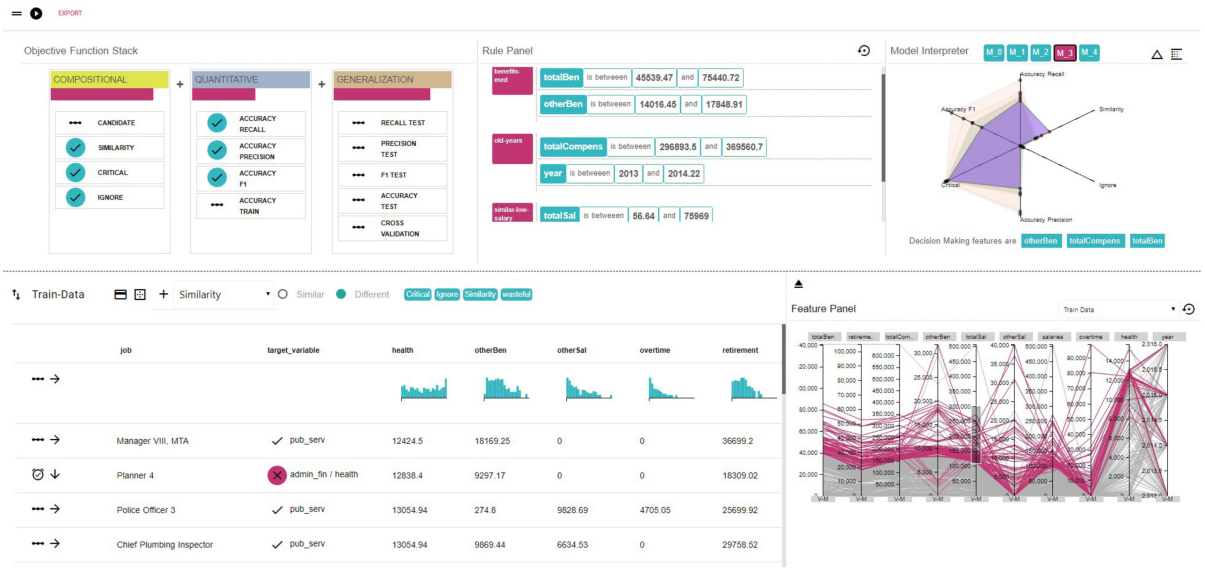

FIGURE 3. QUESTO[8] allows users to interactively specify constraints of objective functions, and observe how well models are able to meet each of them (see Claim 4).

driven encoding. Poor design choices or focusing on the wrong task can easily invalidate visualization as a trust-building tool for AI, no matter if by accident or by agenda. The human–computer interaction (HCI) and visualization (Vis) community should establish a workable theory of how and if trust can be built toward a system at all, or only between humans using the tools and humans creating them.

Equally important is the clear understanding of the AI models themselves and the limitations of "explainable artificial intelligence" (XAI) techniques. For instance, if visualization only explains parts of a model or local decisions, it might introduce confirmation bias or unjustified trust in a model. Future research should build effective bridges between AI and humans via a thorough understanding of where visualization is truly beneficial to increase the trust in AI.

Visualization is Crucial for Human Acency in AI

While we argue that visualization is not equipped to address all aspects of trust yet, it is a core enabling technique to do so. Interactive visualization is the interface between humans, algorithms, and data; thus, it serves an utmost important role in fostering trust in AI methods. Mentioned by the EU Guidelines for Trustworthy AI,[6] the need to provide transparency—including traceability, explainability, and communication—reinforces the important role that visualization can serve.

Chen *et al.*[9] addressed the broad challenge of what visualization is really for, with saving time as one of their core arguments. The advent of AI has made predominant other areas where visualization is a critical enabling facility to create a bridge in human–machine partnership, especially in the context of trust.

In their encompassing survey, Chatzimparmpas *et al.*[10] identified many visualization techniques which help in understanding parts of the AI process, e.g., AI model visualization, parameter space visualization, or data and prediction uncertainties. These are already available to foster trust in AI for experts. Recent work also studies adaptive visual interfaces. These results propose smart layouts and visual representations, and hence can compensate for or augment user abilities, taking into account individual visual literacy, domain knowledge, expectations, and goals. Furthermore, when adaptation is dynamic, it can account for and leverage learning, changes in preferences, values, and evolution of mental models. Putting such methods into practice can enable trust and agency also for nonexperts, such as end-users and other stakeholders, besides AI experts. Also, on an organizational level, visualization can play a key role in enabling integration of information at different levels of granularity, from different perspectives to expose multifaceted aspects of the data to inform policy makers or support legal disputes.

These are very promising perspectives for supporting trust in AI by visualizing certain aspects of AI-based data analysis, and for different tasks and users.

How to integrate and extend these methods, to support the AI process end-to-end and for all stakeholders and users, and for all modes of involvement and interaction, remains an important challenge for the visualization field. For example, one line of research is in probing the design space of what the appropriate medium is to foster communication between people and AI-driven systems. This benefits people byfostering trust through being able to observe the model outputs visually and inspect specific model parameters directly. Additionally, when domain-specific inaccuracies or discrepancies are found in the outputs of these models, user feedback is often used to incrementally adapt the models to be better-suited to specific domains or decisions. Recently, Das *et al.*[8] presented the concept of interactive objective functions as the method for this user feedback to be incorporated. Their approach allows users to specify particular aspects of the model and data that are important for the task or domain, which are then translated into constraints of the objective function used to generate and select models through the use of AutoML techniques. Figure 3 shows the visual interface for users to adjust constraints and observe how well models are able to meet each of them. Additional such research can pinpoint other areas where providing people interactive visual representations of parameters or structural components of AI models can help promote trust.

CONCLUSION

AI continues to make an impact on a myriad of data-rich application areas. As this trend continues, there is a growing need for tools that help people gain a better understanding of these technologies. People from a variety of backgrounds are faced with trying to gain a holistic understanding of AI's trustworthiness, fairness, interpretability, accountability, and other related factors. In response, this manifesto discusses the role of interactive visualization as a method to promote and foster such concepts. We highlight four claims and discuss their rationale. Further, we indicate areas where interdisciplinary teams, research, and methodologies are required to make an impact. Through this manifesto, we aim to inspire future research that will result in impactful outcomes toward the goal of achieving informed trust in AI.

ACKNOWLEDGMENTS

This Viewpoint is the result of discussion at Dagstuhl Seminar on Interactive Visualization for Fostering Trust in AI (seminar 22351).

REFERENCES

1. F. Sperrle, M. El-Assady, G. Guo, D. H. Chau, A. Endert, and D. Keim, "Should we trust (x) AI? Design dimensions for structured experimental evaluations," 2020, *arXiv:2009.06433*.
2. P. Wright *et al.*, "A comparative analysis of industry human-AI interaction guidelines," 2020, *arXiv:2009.06433*.
3. S. Wickramasinghe, D. L. Marino, J. Grandio, and M. Manic, "Trustworthy AI development guidelines for human system interaction," in *Proc. 13th Int. Conf. Hum. Syst. Interaction*, 2020, pp. 130–136.
4. S. Coppers *et al.*, "Intellingo: An intelligible translation environment," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–13.
5. A. Smith-Renner *et al.*, "No explainability without accountability: An empirical study of explanations and feedback in interactive ML," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–13.
6. H.-L. E. G. on AI, "Ethics guidelines for trustworthy AI." Accessed: 2021. [Online]. Available: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf
7. W. Han and H. J. Schulz, "Beyond trust building—Calibrating trust in visual analytics," in *Proc. IEEE Workshop TRust EXpertise Vis. Analytics*, 2020, pp. 9–15.
8. S. Das, S. Xu, M. Gleicher, R. Chang, and A. Endert, "Questo: Interactive construction of objective functions for classification tasks," *Comput. Graphics Forum*, vol. 39, no. 3, pp. 153–165, 2020.
9. M. Chen, L. Floridi, and R. Borgo, "What is visualization really for?," in *The Philosophy of Information Quality*. Cham, Switzerland: Springer, 2014, pp. 75–93.
10. A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren, "The state of the art in enhancing trust in machine learning models with the use of visualizations," *Comput. Graphics Forum*, vol. 39, no. 3, pp. 713–756, 2020. [Online]. Available: https://doi.org/10.1111/cgf.14034
11. F. Sperrle, H. Schäfer, D. Keim, and M. El-Assady, "Learning contextualized user preferences for co-adaptive guidance in mixed-initiative topic model refinement," *Comput. Graphics Forum*, vol. 40, pp. 215–226, 2021. [Online]. Available: https://doi.org/10.1111/cgf.14301

12. M. J. Krocak *et al.*, "Thinking outside the polygon: A study of tornado warning perception outside of warning polygon bounds," *Nature Hazards*, vol. 102, pp. 1351–1368, 2020. [Online]. Available: https://doi.org/10.1007/s11069-020-03970-5

EMMA BEAUXIS-AUSSALET is an Assistant Professor of ethical computing at Vrije Universiteit Amsterdam, User-Centric Data Science Group, and Lab Manager of the Civic AI Lab. She is the corresponding author of this article. Contact her at e.m.a.l.beauxisaussalet@vu.nl.

MICHAEL BEHRISCH is an Assistant Professor of visual analytics at the Visualization and Graphics Group, Department of Information and Computing Sciences, Utrecht University. Contact him at m.behrisch@uu.nl.

RITA BORGO is a Senior Lecturer in data visualization and the Head of the Human Centred Computing Group, King's College London, London, U.K. Contact her at rita.borgo@kclac.uk.

DUEN HORNG CHAU is an Associate Professor of computing at Georgia Tech, working at the intersection of visualization and machine learning. Contact him at polo@gatech.edu.

CHRISTOPHER COLLINS is an Associate Professor of computer science and Canada Research Chair in Linguistic Information Visualization at Ontario Tech University. Contact him at christopher.collins@ontariotechu.ca.

DAVID EBERT is the Gallogly Professor of Electrical Engineering and Computer Science, an Associate Vice President of Research and Partnerships, and the Director of the Data Institute for Societal Challenges at the University of Oklahoma. Contact him at ebert@ou.edu.

MENNATALLAH EL-ASSADY is a Research Associate with the University of Konstanz. Contact her at menna.el-assady@uni.kn.

ALEX ENDERT is an Associate Professor with the School of Interactive Computing, Georgia Tech. Contact him at endert@gatech.edu.

DANIEL A. KEIM is the Head of the Data Analysis and Visualization Research Group, and a Professor of computer science at the University of Konstanz, Germany. Contact him at keim@uni-konstanz.de.

JÖRN KOHLHAMMER is the Head of the Competence Center for Information Visualization and Visual Analytics at Fraunhofer IGD, and Honorary Professor of user-centered visual analytics at TU Darmstadt, Germany. Contact him at joern.kohlhammer@igd.fraunhofer.de.

DANIELA OELKE is a Professor of machine learning at the Offenburg University, Germany. Contact her at daniela.oelke@hs-offenburg.de.

JAAKKO PELTONEN is a Professor of statistics and data analysis at the Tampere University, Faculty of Information Technology and Communication Sciences, and head of the Statistical Machine Learning and Exploratory Data Analysis research group. Contact him at jaakko.peltonen@tuni.fi.

MARIA RIVEIRO is a Professor of computer science at the School of Engineering, Jönköping University, Sweden, working on machine learning and HCI aspects. Contact her at maria.riveiro@ju.se.

TOBIAS SCHRECK is a Professor and the Head of the Institute of Computer Graphics and Knowledge Visualization, Graz University of Technology, Austria. Contact him at tobias.schreck@cgv.tugraz.at.

HENDRIK STROBELT is Research Scientist at IBM Research and Explainability Lead at the MIT-IBM Watson AI Lab. Contact him at strobelt@mit.edu.

JARKE J. VAN WIJK is a Professor in visualization with the Department of Mathematics and Computer Science at the Eindhoven University of Technology. Contact him at j.j.v.wijk@tue.nl.

Contact department editor Theresa-Marie Rhyne at theresamarierhyne@gmail.com.