Human Trust Modeling for Bias Mitigation in Artificial Intelligence

Fabian Sperrle
Udo Schlegel
fabian.sperrle@uni.kn
u.schlegel@uni.kn
University of Konstanz

Mennatallah El-Assady
Daniel Keim
mennatallah.el-assady@uni.kn
daniel.keim@uni.kn
University of Konstanz

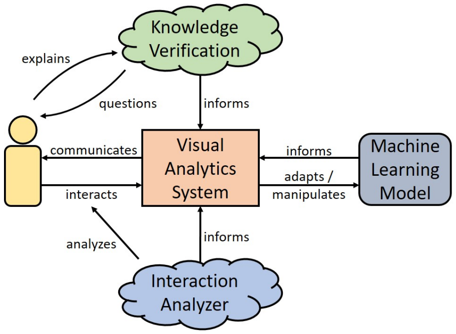


Figure 1: Human Trust Modeling can be integrated into Visual Analytics systems through the addition of *Interaction Analysis* and a *Knowledge Verification* Loop.

A human-in-the-loop Visual Analytics system as shown in Figure 1 is characterized through interaction loops between the user, the interface, and the underlying model. To enable Human Trust Modeling, those interactions need to be analyzed and modeled. Knowledge verification loops provide calibration data that allows the system to judge whether the user has understood the underlying modeling processes.

POSITION STATEMENT

Human involvement in machine-learning processes is often necessary due to the limitations of current models and the non-availability of domain knowledge to those systems. This involvement introduces the problems of bias and model manipulation. Systems should model their trust-level towards their users. Drastic model changes should not be allowed until users have proven that they are consistent in their interactions.

ABSTRACT

Human-in-the-loop model-building processes are increasingly popular as they incorporate human intuition and not easily externalized domain knowledge. However, we argue that the inclusion of the human, and in particular direct model manipulations, result in a high risk of creating biased models and results. We present a new approach of “Human Trust Modeling” that lets systems model the users’ intentions and deduce whether they have understood the underlying modeling processes and act coherently. Using this trust model, systems can enable or disable and encourage or discourage interactions to mitigate bias.

INTRODUCTION AND BACKGROUND

Modern machine learning systems suffer from several issues. They usually require large sets of training data and cannot utilize human intuition – they are often black box models, excluding the human. Those blackbox models are not satisfactory: they lead to unexpected results, do not promote understanding or trust, and are susceptible to corrupt data.

An attempt to mitigate the shortcomings of those traditional models has been the inclusion of “the human in the loop”, providing their domain knowledge whenever necessary. This has inspired several fields of work like active learning and various mixed-initiative approaches. While humans are slower at many tasks than machines and cannot analyze high-dimensional spaces as effectively, they can often spot wrong modeling directions early, and correct them.

A common argument postulates that the inclusion of human reasoning in the modeling process can help to prevent bias and overcome issues created by limited, biased or wrong data. Endert et al. go one step further and demand “we must move beyond human-in-the-loop” to “human **is** the loop.” [3]

HUMAN BIAS IN VISUAL ANALYTICS

Wall et al. [8] have identified four perspectives on human bias in Visual Analytics:
Cognitive Processing Error: Includes cognitive biases like anchoring bias, confirmation bias, or the attraction effect.
Filter for Information: "How sensory information is distinguished and interpreted".
Preconception: (Unconscious) biases formed through the users domain knowledge, previous experience, and expectations.
Model Mechanism: Bias as defined in cognitive modeling is used to determine if users are biased towards certain data points or dimensions.

LEARNING WITHOUT THE HUMAN

AlphaGo [5] and now AlphaZero [4] have shown that it is possible to train an AI by self-play and established a new thesis: "*A decision problem for intelligence is tractable by a human being with limited resources if and only if it is tractable by an AlphaGo-like program.*" [9] This thesis establishes a new era of AI. Data to train is generated by the machine and is only tested against the task definition. The outcome is also tested against the task definition and needs a quality metric or measure to be defined.

EXPLAINABLE AI (XAI)

Both during and after model training XAI highlights decision steps that led to a given model outcome [1]. Such explanations help humans to build trust in AI, as they make some of the inner workings understandable. The general idea is to make an AI more human-like [1] by enabling it to explain itself and its decisions.

Their thesis is "that a little domain knowledge goes a long way" [3]; potentially further than a more powerful algorithm. This inclusion of humans into the core of the processing loop gives the users great power over the analysis process and its results. At the same time, "little consideration has yet been given to the ways inherent human biases might shape the visual analytics process." [7] Wall et al. have identified four different views on human bias in Visual Analytics [8]. In this paper we focus on those human biases, rather than biases created by skewed training data, for example.

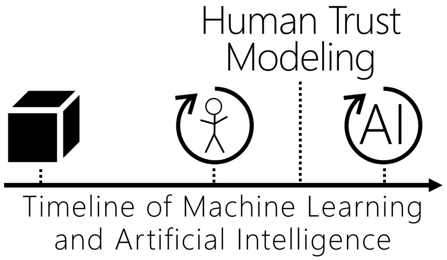
Machine learning developed into using blackbox models without human feedback loops and is currently transitioning towards more human involvement. In this position paper, we argue that human-in-the-loop processing cannot be the solution to current machine learning problems, but is rather an intermediate step on a pathway towards more automated, trustworthy and explainable Artificial Intelligence, with less human involvement. We contribute the novel concept of Human Trust Modeling from an AI perspective as a technique to avoid bias in current systems.

ADVANTAGES OF KEEPING THE HUMAN OUT OF THE LOOP

The definition of AI describes a machine which can perform tasks as well or better than a human. The benefits of moving the human out of the loop are based on this theoretical ability. AI-based methods reached or improved the state-of-the-art for various tasks, and can even exceed human performance for some while being faster and thus more efficient.

The AlphaGo thesis (see *Learning without the Human*) introduces a new direction and shows how it is possible to move the human out of the loop to achieve state-of-the-art efficiency. According to this thesis, systems can generate their training data by self-play and consequently, remove the human from the loop. Self-play or self-learning enables the AI to train outside of the limitations of human understanding and data and develops solution strategies that humans would not contrive. It, therefore, develops a basic domain knowledge without human involvement and reduces the areas in which humans can introduce bias. Together with a faster decision-making process, this can lead to AI exceeding human performance. While its application to general, underspecified AI problems remains an open challenge, self-learning is effective for constraint scenarios with clear task definitions and well-defined, computationally tractable quality metrics [9], as showcased for Go [5] and Chess [4].

Self-learning minimizes the human bias, but also reduces the understandability of the results and the trust of the human towards the AI. But how can we look into the decision process and understand the pattern the machine recognized in the data, the domain, or the game? How can we understand the training process of the AI to learn from it and rebuild trust? To answer these questions, explainable AI (XAI) is needed [2]. Good explanations of the underlying AI models allow the user to understand them from "outside of the loop". In combination with self-learning, XAI can be used to extract novel domain knowledge from an AI. Consequently, user involvement in modeling processes can be reduced to deciding whether results are correct or not.

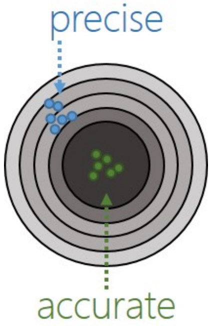


PRECISION AND ACCURACY

In the context of Human Trust Modeling, we understand those terms as follows:

Precision. The coherence between all interactions of one user.

Accuracy. The "correctness" of all interactions of one user with respect to the training data and the interactions of the crowd.



We expect non-biased users to be both precise and accurate. However, accuracy is difficult to determine in the presence of potenially biased training data.

HUMAN-TRUST-MODELING

As outlined in the previous section, many benefits arise from removing humans from modeling loops. However, this is not a valid way forward at the current stage of AI research, as the underlying systems are often not yet efficient and powerful enough to produce meaningful results, and reverting to black boxes does not help. Instead, modern human-in-the-loop systems need to develop from detecting and quantifying bias to preventing it.

We propose to model the user from an AI perspective as shown in Figure 1. Depending on their expertise, each user is assigned an initial trust score. At the beginning of the analysis, the user is a blackbox for the AI. As the analysis progresses, every interaction paints a more detailed picture of the user to the AI and informs a trust-calibration process. Here, the AI can, for example, deduce whether the user's interactions are precise. Is the user blindly trying to change things? Is a pattern recognizable? Is the user checking the provided on-demand-details before initiating modifications? Trust scores should increase when the shown behavior is precise, and decrease otherwise.

A weaker influence on the trust level is given by the accuracy of performed interactions with respect to trends and patterns in the training data. This influence is intentionally lower as completely unbiased training data and is practically impossible to obtain for most (complex) tasks. Instead of comparing patterns to those in the training data, systems can also compare user interactions to those of a crowd. If no crowd is available, it could be generated by virtual AI agents. While not necessary for the trust model or the analysis process in general, trust levels could be increased further if users are able to demonstrate their understanding of the underlying modeling processes. Such understanding can be demonstrated by answering verification or transfer questions, or rating proposed model changes as precise or imprecise. We do, however, emphasize that understanding of the modeling processes does not prevent a user from intentionally or unintentionally introducing bias into a system, making understanding of the modeling processes on its own not particularly useful as a metric.

The current trust level of the user as seen by the AI should directly inform which interaction possibilities are encouraged or discouraged. In extreme cases, disabling interactions might be necessary. However, it is unlikely that machines can decisively overrule expert users with specific domain knowledge in the near future. Instead, systems should focus on promoting exploration of other data points, educate the user about potential correlations between their interactions and bias, or actively propose alternative interactions to consider. Such alternative interactions could be generated using concepts like *Speculative Execution* [6], ranked according to some quality metric(s), and presented to the user with a request for feedback.

Using Human Trust Modeling, systems can enter a more controlled collaboration with the user. Bidirectional trust calibration enables more efficient human-in-the-loop model optimization and reduces the risk of bias-introductions in that process.

RESEARCH OPPORTUNITIES

- Which interaction types or sequences are most likely to introduce bias?
- How can malicious users be kept from exploiting the system by deliberately being precise but inaccurate?
- What are the rules and processes of crowd-based trust calibration?
- How can we model trust propagation to reflect (evolving) user expertise levels?

TAKE-HOME MESSAGES

- Research should not only investigate how to measure bias, but also how to prevent this bias.
- Human Trust Modeling reverses well-known trust-building theory to model the user from an AI perspective.
- Bidirectional trust between user and AI enables better collaboration and more efficient model optimization.

CONCLUSION AND RESEARCH OPPORTUNITIES

We realize that the proposed system has high user-frustration potential. To avoid abandonment, changes in the trust level, as well as the current score, need to be clearly communicated to users together with an explanation. If users lose trust in the AIs ability to compute a trust score correctly, acceptance for limited interaction possibilities can be expected to decrease rapidly.

Future research should investigate which interactions should be disabled or limited by what factor to minimize user frustration while remaining effective at preventing bias. It will also be essential to detect malicious users that prove their understanding of the system coherently, just to intentionally insert bias in the next step. Here, approaches using different human and AI agents seem most promising in the absence of quantifiable, externalized domain knowledge bases.

We have presented human-trust modeling from an AI perspective, a novel approach towards debiasing human-in-the-loop systems. We have identified open research questions leading towards the practical applicability of Human Trust Modeling as a bias-prevention technique.

REFERENCES

[1] David Alvarez-melis and Tommi S Jaakkola. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. NeurIPS (2018). arXiv:arXiv:1806.07538v2

[2] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. Ml (2017), 1–13. https://doi.org/10.1016/j.bbrc.2004.04.155 arXiv:1702.08608

[3] Alex Endert, M. Shahriar Hossain, Naren Ramakrishnan, Chris North, Patrick Fiaux, and Christopher Andrews. 2014. The human is the loop: new directions for visual analytics. *Journal of Intelligent Information Systems* 43, 3 (12 2014), 411–435. https://doi.org/10.1007/s10844-014-0304-9

[4] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. 2017. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. (2017), 1–19. https://doi.org/10.1002/acn3.501 arXiv:1712.01815

[5] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George Van Den Driessche, Thore Graepel, and Demis Hassabis. 2017. Mastering the game of Go without human knowledge. *Nature* 550, 7676 (2017), 354–359. https://doi.org/10.1038/nature24270 arXiv:1610.00633

[6] Fabian Sperrle, Jürgen Bernard, Michael Sedlmair, Daniel Keim, and Mennatallah El-Assady. 2018. Speculative Execution for Guided Visual Analytics. In *Workshop for Machine Learning from User Interaction for Visualization and Analytics at VIS 2018*. 1–5.

[7] Emily Wall, Leslie M. Blaha, Lyndsey Franklin, and Alex Endert. 2017. Warning, Bias May Occur: A Proposed Approach to Detecting CognitiveBias in Interactive Visual Analytics. In *IEEE Conf. on Visual Analytics Science and Technology*. 1–12.

[8] Emily Wall, Leslie M. Blaha, Celeste Lyn Paul, Kristin Cook, and Alex Endert. 2018. *Cognitive Biases in Visualizations*. Springer International Publishing, Cham, Chapter Four Perspectives on Human Bias in Visual Analytics, 29–42.

[9] Fei Yue Wang, Jun Jason Zhang, Xinhu Zheng, Xiao Wang, Yong Yuan, Xiaoxiao Dai, Jie Zhang, and Liuqing Yang. 2016. Where does AlphaGo go: From church-turing thesis to AlphaGo thesis and beyond. *IEEE/CAA J. Autom. Sin.* 3, 2 (2016), 113–120. https://doi.org/10.1109/JAS.2016.7471613