Effect of Transparency and Trust on Acceptance of Automatic Online Comment Moderation Systems

Jens Brunk*, Jana Mattern†, Dennis M. Riehle‡

European Research Center for Information Systems (ERCIS)
University of Münster, Leonardo-Campus 3, 48149 Münster, Germany

*jens.brunk@ercis.uni-muenster.de, †jana.mattern@wiwi.uni-muenster.de, ‡dennis.riehle@ercis.uni-muenster.de

***Abstract*—User-generated online comments and posts increasingly contain abusive content that needs moderation from an ethical but also legislative perspective. The amount of comments and the need for moderation in our digital world often overpower the capacity of manual moderation. To remedy this, platforms often adopt semi-automated moderation systems. However, because such systems are typically black boxes, user trust in and acceptance of the system is not easily achieved, as black box systems can be perceived as nontransparent and moderating user comments is easily associated with censorship. Therefore, we investigate the relationship of system transparency through explanations, user trust and system acceptance with an online experiment. Our results show that the transparency of an automatic online comment moderation system is a prerequisite for user trust in the system. However, the objective transparency of the moderation system does not influence the user's acceptance.**

***Index Terms*—transparency, trust, acceptance, automatic, comment-moderation, user posts**

I. The Need for Online Comment Moderation Systems

User-generated online comments and posts have changed. If comments[1] were constructive and respectful in former days, it is not like that anymore. In an attention-grabbing article from 2016, The Guardian describes this change as "the rising global phenomenon of online harassment". Regularly receiving comments that include "xenophobia, racism, sexism and homophobia", the authors describe this phenomenon as "the dark side of Guardian comments" [1]. Not surprisingly, the detection of abusive content in user-generated online comments has become an important issue for many stakeholders [2]. The political establishment is aware of the issue as well. In Germany, where social media sites displayed many hateful comments against refugees in the last years, the authorities formed a task force with the goal of enforcing providers to filter user comments for hateful content [3]. While this was also criticized for being censorship, similar efforts were conducted in other European countries. In consequence, several newspapers and social media platforms in Europe closed down their comment section [4], making it impossible for users to discuss current topics.

[1]We refer to both user posts and user comments as comments here, since they are mostly posted in reaction to other content.

Online comment moderation systems are an approach to prevent closing comment sections. While many of us primarily see Facebook, Twitter and Google as leading digital media platforms, there are many other websites, which allow users to leave comments. Such websites usually do not have the resources to build analysis tools by themselves and could greatly benefit from the existence of an online comment moderation system. Such systems filter incoming comments based on a suitable methodology and may either publish or reject a comment or put an incoming comment in a queue for manual moderation. The automatic detection of abusive content in user-generated comments is nothing new. Several researchers have tried to tackle this challenging problem [5]. For example, Warner and Hirschberg [6] use word-sense disambiguation to identify hate speech in websites, Burnap and Williams [7] detect hate speech in Twitter posts using a method called "bag of words", where natural language is tokenized using stopword lists and stemming, and Waseem and Hovy [8] analyze tweets using "n-grams", where text input is converted into tokens of length n (usually 2 or 3) taking into account the ordering of words in a sentence. Slightly more advanced is the work of Nobata et al. [2] who combine the "bag of words" and "n-grams" approaches with deep-learning technologies. Köffer et al. [9] applied and compared the previous approaches in the context of the German refugee crisis in 2016.

Methods for detecting abusive content have one thing in common: They provide a decision model, which separates acceptable from non-acceptable comments. The degree of comprehensibility and traceability of such a decision model may differ largely between the approaches. While decision trees are rather easy to understand for humans, more complex mixed-method approaches based on deep-learning lead to better results [2]. Due to the high number of parameters that do not have a semantic meaning, such approaches are harder to understand and trace. Recently, a new European Union (EU) regulation on algorithmic transparency was introduced. This regulation states that individuals have a right for an explanation of the decisions made [10]. Consequently, due to the complex nature of the above mentioned methods, this calls for a higher transparency of the decision models that interact with the user generated comment [11], [12]. Transparent decisions are as well desirable from an economic perspective, since non-

understandable decisions may lead to an aversion against computational intelligence [13], hence hindering the acceptance of an online comment moderation system. Acceptance of an intelligent system is a complex construct to study and influencing factors can be manifold. Given the fact that machine learning based decisions are hard to understand and users therefore often oppose automatic decisions, we try to examine the role of trust and transparency for users' acceptance of such a system. Our research question is therefore: How do the perceived transparency and trustworthiness of a moderation system influence its acceptance? The remainder of this paper is structured as follows: First, we describe our model foundations and related work. Next, we describe the methodology and results of our empirical study. Finally, our paper closes with the discussion of our results and an outlook on future work.

II. Related Work and Model Foundations

This section introduces the key constructs of explanations, transparency, trust and technology acceptance as they are used in our model and embeds it in the relevant related literature.

Explanations. Nowadays, many decision support systems are created that base their decisions on large amounts of data. To evaluate all this data, computational methods, such as artificial intelligence or machine learning, are applied. Users often perceive these methods as black boxes, since they hide their internal processes and decision logic. Once a system affects the user, they want to understand how it works. A system provider can try to reach this understanding by explaining the system and its decisions. Many years ago, Gregor and Benbasat laid the ground-work for the theoretical understanding of explanations for intelligent systems [14]. They provide an overview of explanatory constructs used in empirical research. In our case, the derived constructs, which describe the outcome of explanation use, i.e., the effect on confidence/trust in judgments or agreement with conclusions, are specifically relevant. One of the propositions that the authors derived was that "explanations will be used when the user experiences an expectation failure, or perceives an anomaly" [14, p. 506], such as an automatic moderation system denying the users comment.

Triggered by the above mentioned EU regulation on algorithmic transparency, Guidotti et al. recently surveyed a large number of methods to explain black box models [15]. They derived a classification of the main problems with respect to explanations and the type of explanations for black box systems.

Transparency. The goal of explanations is for the user to understand the system. This understanding leads to the perception of transparency. According to Cramer et al., "transparency aims to increase understanding and entails offering the user insight in how a system works, for example by offering explanations for system behavior" [16]. We follow their concept of transparency for the remainder of this work.

Intelligent systems, as Gregor and Benbasat call them, are applied in many areas, such as knowledge based systems [17], recommender systems [18] or algorithmic media [19]. All systems use explanations for decisions to increase their transparency.

Trust. A secondary objective of the use of explanations for intelligent systems is to gain the users' trust. Trust is a complex and multi-dimensional construct that has been discussed in detail in previous information systems (IS) research [20]–[25]. We pick up on this groundwork by understanding trust as a construct where the user has confidence in the system and trusts that it will "behave capable [...], ethically [...], and fairly" [25, p. 123] towards him. Explaining the system's decisions serves the purpose that the user believes that the system works correctly and is not biased or discriminating against someone. Similar to transparency, trust also has been investigated in many application domains of intelligent systems, such as e-commerce [22] and recommendation agents [26].

Technology Acceptance. Technology acceptance research has been very popular in IS research in the past. Several models exist that try to explain how and why someone uses an IS. The most predominant one is the Technology Acceptance Model (TAM) that was developed by Davis [27]. At its core, the TAM determines technology use and acceptance through perceived usefulness and perceived ease of use, which are influenced by several external factors. The TAM has spawned a great amount of research on these influencing factors, including trust and transparency [24], [25], which extended the original model. The many extensions ultimately lead to the development of a unified theory of acceptance and use of technology (UTAUT) [28]. All of these models are based on the theory of reasoned action (TRA) [29] and theory of planned behavior (TPB) [30]. As pointed out by Benbasat and Barki in their paper "Quo vadis, TAM?" [31] the insular focus on TAM application to different scenarios in IS research has a number of drawbacks and should be avoided. Therefore, we also take a step back, and consider only the very central insight of all these models, that the intention to use an IS is the precedent to its actual utilization by the user.

III. Research Framework

Our research framework connects the previously introduced key constructs of our model and we formulate three underlying hypotheses. In line with previous research [16], we use explanations to increase the perceived transparency. We suggest that alternating the quality of explanation for the system's decision influences the degree of perceived transparency.

Hypothesis 1: A high degree of detail in the explanation for the system's decision increases the perceived transparency.

We previously showed that explanations for transparency and trust are closely connected. In related application domains, it was found that users feel more confident in recommendation agents, if they understand them [32], [33]. Gedikli et al. argue that "user-perceived transparency is also an important factor for trust" [18, p. 379]. We therefore state that

Hypothesis 2: Perceived transparency predicts trust.

Trust is regularly found to be a major factor in the acceptance of intelligent systems [14], [33]–[35]. As such, a higher level of trust is associated with an increased chance of the user intending to adopt the system [20], [22], which leads to our third hypothesis.

Hypothesis 3: Perceived trust predicts the acceptance of a moderation system.

Hypothesis two and three together with the related concepts are displayed in Fig. 1.

Fig. 1: Research Framework

IV. Research Method

A. Experiment Design

To test for and evaluate the relationship between transparency, trust and acceptance, we designed an online experiment. In the experiment, we simulated the act of writing and submitting a comment to a short news article. After submitting a comment, the moderation system informed the user whether the comment was published or needed moderation.

To date, the only mature comment moderation systems that work with computational methods to support platform moderators are Jigsaw's Perspective API (a sub company of Google) and Mozilla's Coral Project, which builds upon the Perspective API. However, these tools and their decision algorithm are not openly available and currently only focus on the English language. As these systems are still at the beginning of their development, we had to simulate the moderation system in our experiment.

We conducted a small pre-study to ensure a realistic simulation. We crawled news articles and comments from different German news websites. To receive an objective evaluation of these comments we used the crowd worker platform Figure Eight[2] to rate 150 of the comments. Each comment was judged by five separate workers for being critical (needs to be moderated) or uncritical (no moderation needed). If a crowd worker selected critical, they then further indicated whether it was critical because of abusive language, insults, hate speech or threats. The categories stem from a consolidation of reporting functionalities and codes of ethics of major social networks and news websites.

For our main experiment, we presented the participants two separate short articles. For each article, we asked them to choose one comment that best represented their attitude from five available comments. We selected the comments based on the averaged criticality evaluation of the five crowd workers, i.e., the participants could choose between a critical, a slightly critical, a moderate, a slightly uncritical and an uncritical comment – without knowing which category each comment stands for. See Fig. 2 for an example of the structure of the article display and comment selection within the experiment. The comments were displayed in a randomized order.

To test the effects of transparency on trust and system acceptance, we randomly distributed the participants to a control and treatment group. The control group represented low transparency. After submitting their comment for the article, the participants only received the information whether their comment was published. The treatment group consisted of a highly transparent moderation system. After submitting the comment, a short text explained why and how the moderation system decided to publish or not publish the comment, including a figure that depicted the rating of the crowd workers, which simulated the system's evaluation (see Fig. 3). The decision of the system was based on the criticality rating that stemmed from the pre-study.

At the end of the experiment, participants evaluated the three constructs of our research framework on a five-point Likert scale (0= completely disagree, 1= rather disagree, 2= undecided, 3= rather agree, 4= agree completely). Items were based on previously validated questionnaires (see the Appendix for more information) and translated into German by a native speaker.

B. Randomization and Attention Checks

We included several measures that ensured the quality of results in online experiments. We included an attention check within the selection of comments for a third news article to filter out respondents that had not properly read the article or comments. Furthermore, we recorded the time spent on each page of the experiment. We randomized the order of the articles (including the attention check), the comments and the items of the construct questionnaire.

C. Participants

We recruited participants through the service of Prolific[3] and implemented the experiment with Qualtrics[4]. We compensated each participant with 1,15£. Since participants needed in average 8 minutes to complete the experiment, our payment equals the German minimum wage and is above the average pay rate on the platform. We additionally offered a bonus payment of 0,15£ for participants who correctly answered a content related question at the end, to motivate them to pay extra attention to the tasks.

We added a pre-screening filter, such that only users who were fluent in German could participate in the study. Then, we ran an a priori power analysis with an alpha of 0.05 and an effect size of d=0.5 to calculate the sample size needed. G*Power [36] recommended an N of 176. The experiment was online for one week, starting the 4th April 2018, resulting in 194 responses.

[2]http://www.figure-eight.com

[3]http://www.prolific.ac

[4]http://www.qualtrics.com

Please read the following article and select the comment that most closely matches your opinion and that you would like to submit as a comment.

<Here is the news article, i.e. actual text. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.>

<Here is a critical comment to select.>

<Here is a slightly critical comment to select.>

<Here is a moderate comment to select.>

<Here is a slightly uncritical comment to select.>

<Here is an uncritical comment to select.>

Fig. 2: Article Display and Comment Selection in the Experiment

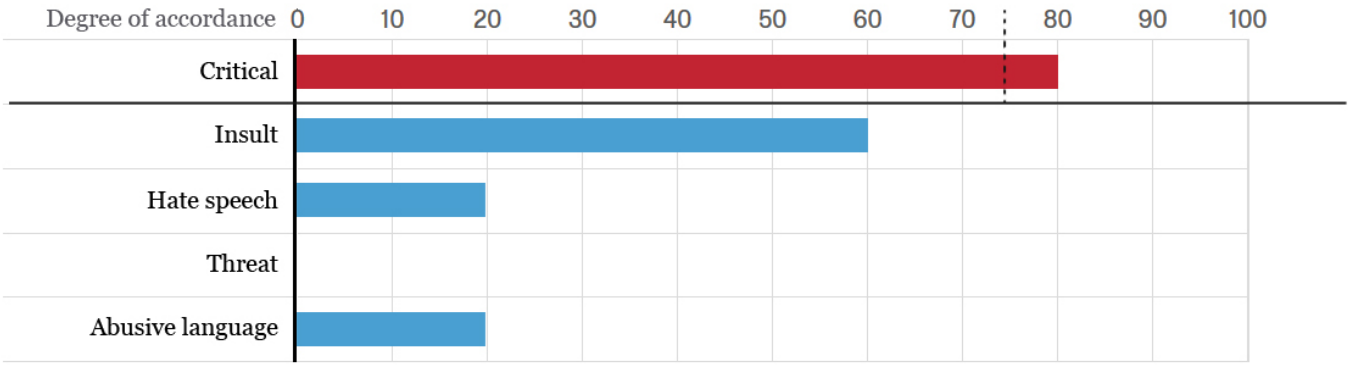

Fig. 3: Simulated Moderation System Result Explanation

To receive valid results, we only considered participants who took an adequate amount of time for reading the text and choosing a comment. Thus, participants who needed less than 100 seconds and more than 2500 seconds for conducting the experiment were excluded from the analyses. Also, 14 participants, who failed the attention check were excluded. The remaining participants still differed to a high amount in the duration for answering the questions (M = 471.48, SD = 275.03). Since individuals differ enormously in how fast and accurately they can read [37] this range is not surprising. Of the remaining 171 participants, 92 were female, 78 male and one participant did not indicate their gender. The treatment group finally consisted of 90 and the control group of 81 participants.

V. Results

A. Descriptive Statistics and Bivariate Correlations

Means (M), standard deviations (SD) and intercorrelations (pearson correlation) of all measures used in the analyses are displayed in Table I. Since we were interested in the overall perceived transparency and trustworthiness, we used the aggregates transparency and trust for the following analyses.

Table II shows descriptive statistics and intercorrelations of the aggregated values (for descriptions of the measurement items see Table III).Results showed that the mean of all variables was close to 2.00, indicating a neutral value. This is not surprising since values of the control as well as of the treatment group were included in the analysis. Moreover, all variables correlated significantly and positively with each other.

B. Hypotheses Testing

We conducted an independent-samples t-test to examine whether the difference in the detail level of the explanation influenced transparency. We found a significant difference in the scores of transparency for the treatment group (M = 2.12, SD = .94) and control group (M = 1.42, SD = .94); t(169) = 4.89, $p < .001$. H1 could therefore be accepted.

To test hypotheses 2 and 3, we calculated a structural equation model by using the R package lavaan. We modeled transparency as a predictor for trust which we expected in turn to predict acceptance.

*** Path coefficients are significant at p < .001 level

Fig. 4: Research Framework

Results confirmed our hypotheses 2 and 3. A high value of transparency leads to a high value of trust ($\beta = 0.55$, p ¡ .01) which in turn increases acceptance ($\beta = 0.49$, p ¡ .001) .

VI. Discussion and Limitations

For our study, we have used two different simulated automatic moderation systems, where one system was designed to be transparent and provide explanations of algorithmic decisions to the user, while the other system was non-transparent and did not provide any explanations. Our results show that a moderation system which gives detailed information regarding the decision of accepting or rejecting a comment, is perceived as more transparent. This suggests that we achieved our goal of creating a more transparent moderation system by explaining the decisions. However, descriptive statistics show that participants were undecided concerning the perceived transparency, trust in the system and acceptance of the system or rather disagreed with the questions. Since the means are calculated using the control group as well as the treatment group, this finding suggests that participants in general have a rather negative attitude towards the moderation system. The high intercorrelations between the variables show that they are positively related.

Furthermore, our results show that participants perceived a transparent system as more trustworthy (hypothesis H2). We also found, that a high trustworthiness is associated to a high probability of acceptance of the system, which is in line with previously mentioned lines of argumentation [20], [22]. It can therefore be suggested, that a transparent system leads to a higher belief in the correctness of the system's decision and therefore to accepting the decision and with that the system itself (hypothesis H3).

Our results imply that a tool which provides a moderation technology must also provide some kind of reasoning for its action in addition to the API that calculates a publish or no-publish decision. Programmatically deriving an explanation for a machine learning inspired decision model is a challenging task, which needs to be covered by future research.

Furthermore, as outlined previously, we have not covered all perspectives of such an automatic moderation system in our experiment. Future research should pick up where we left off, further investigating the user and moderator perspective and how they influence the intention to use and the acceptance of automatic moderation systems. Additionally, we have not covered the aspect of users trying to trick the moderation system. If the system's algorithm is transparent, it might be easier for users to manipulate the system and having bad comments bypass the system. However, as our approach targets a semi-automatic comment system, i.e., a system with human moderators in the loop, we consider this problem to be less relevant, as manipulative behaviour of users would be detected by human moderators and would lead to re-training of the system on new data.

There are some limitations to our approach. Most notably, we do not yet have access to a functioning automatic moderation system or its prototype. Since we could only simulate the act of commenting and comment evaluation, participants had to choose a comment that may not completely represented their own opinion. Furthermore, the constructs of trust, acceptance and adoption are very complex in their nature. In future research, we intend to evaluate these complex effects, our definition and measurements in more detail with longitudinal studies on trust and acceptance in combination with actual system implementation in real scenarios.

Online surveys are a popular research method because they are easy to set up, easy to execute and cost less time and money than regular surveys. However, online surveys and experiments have a number of drawbacks, which do not always make them a good fit. Evans and Mathur evaluated the strengths and weaknesses of online surveys in detail [39]. Fortunately, most weaknesses, such as skewed attributes of Internet population, technological variations, impersonality, privacy issues, etc., do not apply to our case as the environment we intend to replicate is writing comments online.

VII. Conclusion and Outlook

In this paper, we have focused on the research question how the perceived transparency and trustworthiness of a moderation system influences its acceptance. With our work, we support the development of online comment moderation systems, which are crucial for small media web-sites to analyze and manage incoming user comments. Since the development and training of good decision models is a demanding task and smaller, regional newspapers usually do not have their own IT

TABLE I: Descriptive Statistics and Bivariate Correlations

	M	SD	TRA 1	TRA 2	TRU 1	TRU 2	ACC
TRA 1	1.19	1.13	-	.486**	.395**	.403**	.454**
TRA 2	1.67	1.19		-	.519**	.466*	.339**
TRU 1	1.78	1.06			-	.604*	.516**
TRU 2	1.96	1.12				-	.468**
ACC	1.62	1.10					-

Note: Correlations are significant at * = p $<$ .05 and ** = p $<$ .01 level.

TABLE II: Descriptive Statistics and Bivariate Correlations of Aggregated Measures

	M	SD	TRA	TRU	ACC
TRA	1.79	1.00	-	.578***	.458**
TRU	1.87	.96		-	.548**
ACC	1.62	1.10			-

Note: Correlations are significant at * = p $<$ .05, ** = p $<$ .01 and *** = p $<$ 0.001 level.

TABLE III: Measurement Items and Sources

Construct		**Item**	**Source**	
Transparency 1	TRA 1	I understand what the moderation system bases its recommendations on.	Transparency	[16]
Transparency 2	TRA 2	The functionality of the moderation system is transparent.	Self-developed	-
Trust 1	TRU 1	The moderation system is trustworthy.	Trust	[24]
Trust 2	TRU 2	The moderation system is capable to evaluate user comments.	Trust	[25]
Acceptance	ACC	I intend to use a discussion platform that deploys this moderation system.	Intention to Use	[38]

department, we think that such a comment moderation system needs to be provided as a service. For such an "Analytics as a Service" (AaaS) tool it is highly important that its conceptual design fosters a high acceptance in the online community, i.e., with the users. Based on our work, we argue that a moderation system needs to provide a clear and transparent reasoning for why it has accepted or declined an individual comment. This will lead to a high perceived transparency and trustworthiness by the end users. Moreover, research is needed to derive design guidelines for the development of such an AaaS tool. Especially, towards a standardized exchange of both user generated content and algorithmic derived decisions.

ACKNOWLEDGEMENTS

The research leading to these results received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 645751 (RISE_BPM).

Furthermore, we thank Dr. Markus Weinmann and Dr. Alexander Simons who have supported us methodologically in the design of our experiment and through many valuable discussions.

REFERENCES

[1] B. Gardiner, M. Mansfield, I. Anderson, J. Holder, D. Louter, and M. Ulmanu, *The dark side of Guardian comments*, 2016.

[2] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive Language Detection in Online User Content," in *Proceedings of the 25th International Conference on World Wide Web - WWW '16*. New York, New York, USA: ACM Press, 2016, pp. 145–153. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2872427.2883062

[3] A. Faiola, "Germany springs to action over hate speech against migrants," *The Washington Post*, 2016. [Online]. Available: https://www.washingtonpost.com/world/europe/germany-springs-to-action-over-hate-speech-against-migrants/2016/01/06/6031218e-b315-11e5-8abc-d09392edc612_story.html?noredirect=on&utm_term=.b499809c07ae

[4] The Coral Project Community, *Shutting down onsite comments: a comprehensive list of all news*, 2016. [Online]. Available: https://community.coralproject.net/t/shutting-down-onsite-comments-a-comprehensive-list-of-all-news-organisations/347

[5] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 1–10. [Online]. Available: http://www.aclweb.org/anthology/W17-1101

[6] W. Warner and J. Hirschberg, "Detecting Hate Speech on the World Wide Web," in *Proceedings of the Second Workshop on Language in Social Media*. Montreal, Canada: Association for Computational Linguistics, 2012, pp. 19–26. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390374.2390377

[7] P. Burnap and M. L. Williams, "Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making," *Policy & Internet*, vol. 7, no. 2, pp. 223–242, 2015. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84930475861&partnerID=tZOtx3y1

[8] Z. Waseem, "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter," in *Proceedings of the First Workshop on NLP and Computational Social Science*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 138–142. [Online]. Available: http://aclweb.org/anthology/W16-5618

[9] S. Köffer, D. M. Riehle, S. Höhenberger, and J. Becker, "Discussing the Value of Automatic Hate Speech Detection in Online Debates," Leuphana, Germany, 2018.

[10] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation"," *AI Magazine*, vol. 38, no. 3, pp. 50–57, Oct. 2017, arXiv: 1606.08813. [Online]. Available: http://arxiv.org/abs/1606.08813

[11] K. R. Fleischmann and W. A. Wallace, "A covenant with transparency: Opening the black box of models," *Communications of the ACM*, vol. 48, no. 5, pp. 93–97, 2005.

[12] P. Owotoki and F. Mayer-Lindenberg, "Transparency of Computational Intelligence Models," in *Research and Development in Intelligent Systems XXIII*, 2007, pp. 387–392.

[13] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology: General*, vol. 144, no. 1, pp. 114–126, 2015.

[14] S. Gregor and I. Benbasat, "Explanations from Intelligent Systems:

Theoretical Foundations and Implications for Practice," *MIS Quarterly*, vol. 23, no. 4, pp. 497–530, Dec. 1999. [Online]. Available: http://www.jstor.org/stable/249487?origin=crossref

[15] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti, "A Survey Of Methods For Explaining Black Box Models," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 93:1 – 93:42, 2018.

[16] H. Cramer, V. Evers, S. Ramlal, M. v. Someren, L. Rutledge, N. Stash, L. Aroyo, and B. Wielinga, "The effects of transparency on trust in and acceptance of a content-based art recommender," *User Modeling and User-Adapted Interaction*, vol. 18, no. 5, pp. 455–496, Nov. 2008. [Online]. Available: http://link.springer.com/10.1007/s11257-008-9051-3

[17] J. S. Dhaliwal and I. Benbasat, "The Use and Effects of Knowledge-based System Explanations: Theoretical Foundations and a Framework for Empirical Evaluation," *Information Systems Research*, vol. 7, no. 3, pp. 342–362, Sep. 1996. [Online]. Available: http://www.jstor.org/stable/pdf/23010989.pdf

[18] F. Gedikli, D. Jannach, and M. Ge, "How should I explain? A comparison of different explanation types for recommender systems," *International Journal of Human-Computer Studies*, vol. 72, no. 4, pp. 367–382, Apr. 2014. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1071581913002024

[19] J. A. Stark and N. Diakopoulos, "Towards editorial transparency in computational journalism," in *Computation+ Journalism Symposium*, 2016.

[20] D. H. McKnight, V. Choudhury, and C. Kacmar, "The impact of initial consumer trust on intentions to transact with a web site: a trust building model," *The Journal of Strategic Information Systems*, vol. 11, no. 3-4, pp. 297–323, Dec. 2002. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0963868702000203

[21] C. L. Corritore, B. Kracher, and S. Wiedenbeck, "On-line trust: concepts, evolving themes, a model," *International Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 737–758, Jun. 2003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1071581903000417

[22] Gefen, Karahanna, and Straub, "Trust and TAM in Online Shopping: An Integrated Model," *MIS Quarterly*, vol. 27, no. 1, pp. 51–90, 2003. [Online]. Available: http://www.jstor.org/stable/10.2307/30036519

[23] P. A. Pavlou, "Consumer Acceptance of Electronic Commerce: Integrating Trust and Risk with the Technology Acceptance Model," *International Journal of Electronic Commerce*, vol. 7, no. 3, pp. 101–134, 2003.

[24] J. F. George, "The theory of planned behavior and Internet purchasing," *Internet Research*, vol. 14, no. 3, pp. 198–212, Jul. 2004. [Online]. Available: http://www.emeraldinsight.com/doi/10.1108/10662240410542634

[25] P. A. Pavlou and M. Fygenson, "Understanding and Predicting Electronic Commerce Adoption: An Extension of the Theory of Planned Behavior," *MIS Quarterly*, vol. 30, no. 1, pp. 115–143, 2006. [Online]. Available: http://www.jstor.org/stable/10.2307/25148720

[26] W. Wang and I. Benbasat, "Empirical Assessment of Alternative Designs for Enhancing Different Types of Trusting Beliefs in Online Recommendation Agents," *Journal of Management Information Systems*, vol. 33, no. 3, pp. 744–775, Jul. 2016. [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=120040656&lang=de&site=ehost-live

[27] F. Davis, "A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results," Ph.D. dissertation, Massachusetts Institute of Technology, Massachusets, 1985.

[28] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, 2003. [Online]. Available: http://www.jstor.org/stable/10.2307/30036540

[29] M. Fishbein and I. Ajzen, *Belief, attitude, intention, and behavior: an introduction to theory and research*, ser. Addison-Wesley series in social psychology. Addison-Wesley Pub. Co., 1975.

[30] I. Ajzen, *From intentions to actions: a theory of planned behavior*, 1985. [Online]. Available: https://books.google.de/books?id=ueD6nQEACAAJ

[31] I. Benbasat and H. Barki, "Quo vadis TAM?" *Journal of the Association for Information Systems*, vol. 8, no. 4, pp. 211–218, Apr. 2007. [Online]. Available: https://aisel.aisnet.org/jais/vol8/iss4/16/

[32] R. Sinha and K. Swearingen, "The role of transparency in recommender systems," in *CHI'02 extended abstracts on Human factors in computing systems*. ACM, 2002, pp. 830–831.

[33] H. Cramer, B. Wielinga, S. Ramlal, V. Evers, L. Rutledge, N. Stash, L. Aroyo, and B. Wielinga, "The effects of transparency on perceived and actual competence of a content-based recommender," in *Semantic Web User Interaction workshop at CHI, Florence, Italy*, 2008.

[34] W. Pieters, "Explanation and trust: what to tell the user in security and AI?" *Ethics and Information Technology*, vol. 13, no. 1, pp. 53–64, Mar. 2011. [Online]. Available: http://link.springer.com/10.1007/s10676-010-9253-3

[35] J. S. Giboney, S. A. Brown, P. B. Lowry, and J. F. Nunamaker Jr, "User acceptance of knowledge-based system recommendations: Explanations, arguments, and fit," *Decision Support Systems*, vol. 72, pp. 1–10, 2015.

[36] E. Erdfelder, F. Faul, and A. Buchner, "GPOWER: A general power analysis program," *Behavior Research Methods, Instruments, & Computers*, vol. 28, no. 1, pp. 1–11, Mar. 1996. [Online]. Available: http://www.springerlink.com/index/10.3758/BF03203630

[37] M. Daneman, "Individual differences in reading skills." in *Handbook of Reading Research*, ser. 2. New York, NY: Pearson, 1991, pp. 512–538.

[38] V. Venkatesh and F. D. Davis, "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies," *Management Science*, vol. 46, no. 2, pp. 186–204, Feb. 2000. [Online]. Available: http://pubsonline.informs.org/doi/abs/10.1287/mnsc.46.2.186.11926

[39] J. R. Evans and A. Mathur, "The value of online surveys," *Internet Research*, vol. 15, no. 2, pp. 195–219, Apr. 2005. [Online]. Available: http://www.emeraldinsight.com/doi/10.1108/10662240510590360