Progressive Disclosure: When, Why, and How Do Users Want Algorithmic Transparency Information?

AARON SPRINGER, Computer Science, University of California at Santa Cruz, Santa Cruz, CA, USA
STEVE WHITTAKER, Computational Media, University of California at Santa Cruz, Santa Cruz, CA, USA

It is essential that users understand how algorithmic decisions are made, as we increasingly delegate important decisions to intelligent systems. Prior work has often taken a techno-centric approach, focusing on new computational techniques to support transparency. In contrast, this article employs empirical methods to better understand user reactions to transparent systems to motivate user-centric designs for transparent systems. We assess user reactions to transparency feedback in four studies of an emotional analytics system. In Study 1, users anticipated that a transparent system would perform better but unexpectedly retracted this evaluation after experience with the system. Study 2 offers an explanation for this paradox by showing that the benefits of transparency are context dependent. On the one hand, transparency can help users form a model of the underlying algorithm's operation. On the other hand, positive accuracy perceptions may be undermined when transparency reveals algorithmic errors. Study 3 explored real-time reactions to transparency. Results confirmed Study 2, in showing that users are both more likely to consult transparency information and to experience greater system insights when formulating a model of system operation. Study 4 used qualitative methods to explore real-time user reactions to motivate transparency design principles. Results again suggest that users may benefit from initially simplified feedback that hides potential system errors and assists users in building working heuristics about system operation. We use these findings to motivate new progressive disclosure principles for transparency in intelligent systems and discuss theoretical implications.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**; • **Social and professional topics** → *Government technology policy*; • **Human-centered computing → HCI theory, concepts and models**; **Interaction paradigms**; **Empirical studies in HCI**;

Additional Key Words and Phrases: Transparency, intelligibility, intelligent systems, machine learning, emotional analytics, expectation violation, explanation, error, progressive disclosure

ACM Reference format:
Aaron Springer and Steve Whittaker. 2020. Progressive Disclosure: When, Why, and How Do Users Want Algorithmic Transparency Information? *ACM Trans. Interact. Intell. Syst.* 10, 4, Article 29 (October 2020), 32 pages.
https://doi.org/10.1145/3374218

We acknowledge NSF grant IIS-1321102 for financial support.
The reviewing of this article was managed by special issue associate editors Oliver Brdiczka, Polo Chau, Minsuk Kahng, Gaelle Calvary.
Authors' addresses: A. Springer, Computer Science, University of California at Santa Cruz, 1156 High St, Santa Cruz, CA, 95064, USA; email: alspringer@ucsc.edu; S. Whittaker, Computational Media, University of California at Santa Cruz, 1156 High St, Santa Cruz, CA, 95064, USA; email: swhittak@ucsc.edu.

2160-6455/2020/10-ART29 $15.00
https://doi.org/10.1145/3374218

1 INTRODUCTION

Machine learning algorithms underlie the many intelligent systems we routinely use. These systems provide information ranging from route planning to recommendations about criminal parole [3, 8, 10, 64]. As humans with limited time and attention, we increasingly defer responsibility to these systems with little reflection or oversight. Nevertheless, intelligent systems face mounting criticisms about how they make decisions; criticisms that are exacerbated by recent machine learning advances like deep learning that are difficult to explain in human-comprehensible terms. Major public concerns have arisen following demonstrations of bias in algorithmic systems with regards to gender, race, and other characteristics [9, 64, 77, 82]. These issues have led to calls for transparency as a solution to the "unintelligible" algorithms that impair the adoption of intelligent systems [31, 64, 85].

Algorithmic transparency is needed for many reasons. Greater transparency potentially increases end user control and improves acceptance of complex algorithmic systems [31, 44, 57, 85]. It can also promote user learning and insight from complex data, as humans increasingly work with complex inferential systems for analytic purposes [44, 45, 73]. Transparency can also enable oversight by system designers. Without such transparency it may be unclear whether an algorithm is optimizing the intended behavior [35, 52] or whether an algorithm accidentally promotes negative, unintended consequences (e.g., filter bubbles in social media [10, 65]). Given these current concerns, it is increasingly possible that transparency, i.e., "a right to explanation," may become a legal requirement in some contexts [31]. These issues have led some researchers to argue that machine learning must be "interpretable by design" [1] and that transparency is essential for the adoption of intelligent systems, e.g., for medical diagnoses [34, 87].

While such calls for transparency are well motivated, it remains unclear exactly how to enact them in practice. Extensive technical research about operationalizing transparency has emerged in the machine learning community, but no clear consensus has resulted [1, 23, 52–54, 85]. Deciding exactly how to implement transparency is difficult, as there are numerous implementation tradeoffs involving explanation fidelity. Making a complex algorithm understandable to end users can require simplification, which often comes at the cost of reduced accuracy of explanation [46, 47, 53, 54, 73]. For example, methods have been proposed to explain neural network algorithms in terms of more traditional machine learning approaches, but these explanations necessarily present approximations of the original algorithms [47, 52, 53, 58].

Some recent empirical studies attempt to examine the effects of transparency on users. However, these studies reveal puzzling and sometimes contradictory effects. In some settings there are expected benefits: Transparency improves algorithmic perceptions, because users better understand system behavior [43, 44, 51]. But in other circumstances, transparency has other quite paradoxical effects. Transparency may erode confidence in a system, with users trusting it less, because transparency leads them to question the system even when it is correct [1, 13, 50]. Providing system explanations may also undermine user perceptions when users lack the attentional capacity to process complex explanations, for example when executing a demanding task [1, 12, 43, 85]. Overall, these results indicate mixed evidence for the benefits of transparent systems.

These inconsistent results suggest the need for more user-centric research to better understand user reactions to transparency information. They also imply that we have yet to identify the appropriate interaction paradigms to present transparency. Machine learning research communities are forging ahead with foundational research on how to implement transparent systems [46, 47, 54, 56, 58, 85], but studies often stop short of actually testing these systems with users [1, 49, 50]. Evaluation is critical, because, as we have seen, user reactions to transparency show quite contradictory results [12, 43, 50]. Some research suggests that the way we present transparency may account for these contradictory results [28, 43].

Our research seeks to bridge this gap between technical approaches to generating explanations and user-centric evaluations. In four studies, we explore and explain users' direct reactions to a transparent personal informatics system that interprets emotions that users express in written text. The domain of emotional analytics is fruitful for transparency research for several reasons. Emotional analytics is perceived as a key application by many users [16, 17, 18, 39, 90]. And users are the ultimate experts on their own emotions, providing them with ground truth assessments about whether the system is correct. Furthermore, making accurate system predictions about emotions is also difficult, allowing us to examine a key challenge for intelligent systems: how to deal with system error [13, 50, 51, 60].

Overall, then, prior work has often addressed transparency from a technical perspective, asking, "What is possible from an algorithmic standpoint?" rather than "What does the user need?" The current article takes a different approach. Rather than asking how we might support transparency we address user-centric questions of why, when, and how users might need transparency information. We then propose user-centric design principles for the presentation of transparency information.

Across multiple studies, we examine user preferences and reactions to transparency by comparing two versions of a working analytics system that predicts users' emotions from written text. Both versions use the same underlying algorithm. The first, *transparent*, version makes overall predictions about the user's affective state, supplementing overall predictions with incremental visual feedback showing how the algorithm made these judgments. The second *non-transparent* version simply predicts the user's overall affective state, without accompanying transparency feedback. We address the following research questions:

- RQ1: Do users prefer transparent systems? What explains their preferences? (Study 1)
- RQ2: How does transparency influence users' models and expectations about systems? (Study 2 and 3)
- RQ3: When do users want transparency information? (Study 2 and 3)
- RQ4: How might we provide more effective transparency information to meet user needs? (Study 4)

1.1 Contribution

Much recent work on transparency has focused on technical explorations of self-explanatory systems. In contrast, here we take an empirical user-centric approach to better understand how to design transparent systems. Four studies provide novel data concerning user reactions to systems offering transparency information. In Study 1, users anticipated that a transparent system would perform better but retracted this evaluation after experience with the system. Study 2 suggests reasons for this altered preference. On the one hand, transparency can help users develop working models of the underlying algorithm. For others, however, positive accuracy perceptions may be undermined when transparency reveals algorithmic errors. Study 3 indicates when users derive most benefits from transparency information. It shows that users are more receptive to, and derive greater understanding benefits from, transparency information when trying to formulate models of system operation. Study 4 explores potential methods for providing transparency feedback, suggesting that users may benefit from simplified feedback that hides potential system errors and assists users in building working heuristics about system operation. We combine data from these four studies to motivate new user-centric progressive disclosure principles for presenting transparency in intelligent systems and discuss implications for transparency theory.

2 RELATED WORK

2.1 Folk Theories of Algorithms and Algorithmic Omniscience

A wealth of prior work has explored issues surrounding algorithm transparency in the commercial deployments of systems for social media and news curation. Social media feeds are often curated by algorithms that may be invisible to users (e.g., Facebook, Twitter, LinkedIn). At one point, many users were unaware that Facebook newsfeeds filtered the posts that their friends made [26]. These users reacted with surprise and sometimes anger when they were shown the filtered posts that were "missing" from their newsfeed. Later research showed that many Facebook users develop "folk theories" of their social feed [24], which are imprecise heuristics about how the system works, even going so far as to make concrete plans based upon such folk theories. This work also indicated that making the design more transparent allowed users to generate multiple folk theories and more readily compare and contrast them [24, 25].

Other work illustrates issues regarding incorrect folk theories in the domain of intelligent personal informatics systems, showing specific challenges in how users understand such systems. Users can be prone to blindly believe outputs from algorithmic systems, a phenomenon referred to as algorithmic omniscience [26, 39, 78] and automation bias [19, 59]. For example, KnowMe [84] is a program that infers personality traits from a user's posts on social media based on Big Five personality theory. KnowMe users were quick to defer to algorithmic judgments about their own personalities, stating that the algorithm is likely to have greater public credibility than their own personal statements (e.g., "...At the end of the day, that's who the system says I am..."). Similar results were found in Hollis et al. [39], showing that participants expected intelligent personal informatics systems to serve as ground truth for their experiences and even attributed superhuman qualities to these devices, e.g., "...[it] could tell me about an emotion I don't know that I am feeling...." Other experiments indicate the risk of such trust, showing that users may believe even entirely random system outputs as moderately accurate [78]. Similarly, giving users placebo controls over an algorithmic interface shows corroborating results [83]; as users with placebo controls felt more satisfied with their newsfeed. Without a standard of transparency in intelligent systems, it may be easy to deceive end-users about system accuracy or indeed whether they are using a real system; this is a dangerous proposition when applications can be so easily distributed.

2.2 Transparency

There is a long history of studying transparency and intelligibility in automated systems [1, 6, 85, 89]. However, the results are mixed, with often contradictory effects on user perceptions. Many studies indicate that transparency improves user perceptions of the system [22, 49, 51]. Others have shown that interventions that simply show algorithm prediction confidence improve users' system perceptions [4, 13, 81]. In extreme cases, animations that simulate transparency can cause users to be overconfident about systems even when they err [28].

Other studies show less positive effects of transparent systems on user perceptions. Exposing users to a hypothetical transparent system can lead them to question the system, resulting in reduced agreement with the system [50, 51]. Muir and Moray conclude that any hint of error in an automated system will decrease trust [60]. However, the effect may be different for high certainty systems, with transparency promoting higher user agreement. More recent work indicates other effects; explanations of how a system is working may increase trust [13, 22, 44, 89], but overly complex explanations may reduce this [1, 12, 43, 86]. How might we explain these differing observations? One suggestion is that these different effects arise from the *expectation violation* that a user experiences. Expectation violation occurs when the system behaves in a way that a user did not anticipate [43, 78, 79, 80]. Ideally transparency should build user confidence in a system, whether

or not the user is experiencing expectation violation [43, 80]. In other words, transparency should aid users who are experiencing fundamental difficulties in understanding system operation, at the same time enriching the understanding of users who already have operational system models. However, it may be that research communities have yet to find the correct interaction paradigms to achieve this.

Recently, the machine learning community has begun grappling with issues of explainability. This may be due to the rise of more inscrutable methods like deep learning or from legal requirements arising from the European Union's GDPR. Some machine learning models are "inherently understandable" such as linear models and Generalized Additive Models [53, 54, 85]. These models can be "explained" to users simply through the linear contributions of their features. Other algorithms such as deep neural nets and random forests are more inscrutable, making it demanding to explain how input features match to output predictions [52, 54, 85]. Various new approaches aim to make these inscrutable algorithms understandable. These may rely on approximating the inscrutable algorithm through a simple local or linear model that can be explained to the end user [54, 73]. However, even with these efforts to generate "inherently understandable" models, there is no clear consensus how to convey these models to users in an understandable way. Furthermore, mapping an inscrutable algorithm to an "understandable" equivalent necessarily results in approximations and loss of fidelity [85]. Most importantly, many such attempts at transparency have not been directly tested with users or instead rely on simulated user studies [5, 54, 61]. The absence of real user feedback makes it challenging to operationalize transparency in ways that positively impact users [1, 23].

2.3 Explanation and Persuasion Theory

People interact with computers and intelligent systems in ways that mirror how they interact with other people [63, 72]. Given that transparency is essentially an explanation of why an algorithm made a given prediction, we can turn to fields such as psychology and sociology for guidance about operationalizing explanations [57]. These fields have a long history of studying human explanation. One approach is to model causal explanation as a form of conversation that is governed by common-sense rules [37] such as Grice's maxims [32]. In addition, when a communication breakdown occurs and an explanation is needed this is remedied by a phenomenon known as conversational repair. Conversational repair is interactive; participants in the conversation collaborate to achieve mutual understanding. This often happens in a turn-by-turn structure with repeated questions and clarifications that depend on the current understanding of the interaction participants [75]. These theories suggest that we might operationalize transparency in ways that fit human communication and repair strategies.

Additionally, we see parallels between how people interact with intelligent systems and theories of persuasion. The Elaboration Likelihood Model (ELM) is a dual process model of persuasion [67]. The ELM posits that two different processes are engaged when a person evaluates an argument, similarly to Kahneman's conception of System 1 and System 2 thinking [41]. In ELM, the central processing route is invoked for high stakes decisions. Central processing is cognitively intensive, involving careful weighing of the argument and complex integration into a person's beliefs. In contrast the peripheral route is invoked when people have lower investment in a decision. Peripheral routes employ mental shortcuts, exploiting heuristic cues such as the status and attractiveness of the speaker, the person's current affect, the number and length of the arguments, and other cues not directly related to the content of the argument. Prior work on intelligent systems seems to align with this dual process model [43]; people understand systems through peripheral routes if their expectations are met, only engaging in central processing when their expectations

are violated [1]. This is also demonstrated in the context of Google search suggestions, where some users feel the cost of processing explanations outweighs their benefits [12].

Explanation and persuasion theory therefore offer human-centric accounts of how users might understand complex systems differently in high- and low-stakes situations. Furthermore, these theories imply that transparency needs to be operationalized in multiple ways that both allow users to understand transparency through both cursory heuristic routes and also through focused effort.

2.4 Emotional Analytics

In the current article, our focus is on how users interact with intelligent personal informatics (PI) systems, which are being increasingly deployed within commercial [16, 39, 68] and research domains [7, 27, 38, 55, 90]. PI systems track how a user behaves on some dimension, whether physical or mental, to help users better understand that behavior. Systems may suggest improvements to this behavior through customized feedback and recommendations [38, 70]. Such personally relevant data potentially allows users to analyze and modify their behaviors to promote well-being [17, 18, 38, 39, 40]. For example, fitness trackers allow people to log and adapt their exercise regime to meet specific health goals. In the current study, we explored user reactions to textual emotional analytics, where an algorithm interprets the emotional characteristics of a reported personal experience. Other work has shown that emotional analytics is evaluated by users as one of the key applications of PI technology [16, 39].

Emotional analytics is a fruitful domain for transparency research for other reasons too. First, it allows users to directly engage with personally relevant data. Other transparency work has often asked users to evaluate algorithms in hypothetical scenarios where participants read about or watch simulations of algorithmic deployments and decisions [28, 50, 51]. Instead our aim was to have users directly experience the algorithm, evaluating it *in situ,* as it made decisions about their own personally generated data [38, 39]. A further critical aspect of emotional interpretation is that users are knowledgeable about their own feelings and experiences, allowing them to directly compare algorithmic interpretations with their own personal evaluations of those emotional experiences. This contrasts with other applications of smart algorithms, such as medical diagnostics, where the end user may not be a domain expert. In these contexts, users might be less able to evaluate the results of algorithmic interpretations [34, 46]. In addition, emotion is highly variable between individuals and previous research demonstrates difficulty in accurately predicting emotion from text [48, 71]. One of the major challenges with intelligent systems is handling errors [50, 51, 60, 81]; explaining algorithmic prediction in this difficult domain of emotional analytics allows us to better understand how users make sense of output that contains errors.

3 RESEARCH SYSTEM: E-METER

We designed a working research platform called the E-meter that tests users' reactions to a transparent system that actively interprets their own affective data. The E-meter system and protocol have been successfully deployed in multiple previous experiments [78, 80]. The E-meter predicts the emotional valence of a user's written personal experience and is described to users as an "algorithm that assesses the positivity/negativity of [their] writing." The E-meter is depicted in Figures 1 and 2, showing non-transparent and transparent versions of the same algorithm. It was built as a client-side web application using HTML5 and Javascript, technologies that give us flexibility in deploying to various populations including in-lab studies and Amazon Mechanical Turk experiments.

There are a number of tradeoffs that need to be made when testing user reactions to transparency. In contrast to many previous hypothetical scenario-based studies, where users are presented with descriptions of a system along with its outputs [50, 51], we instead chose to deploy a

Fig. 1. E-meter Non-Transparent Feedback: Meter bar position and color show an overall negative system evaluation of emotions expressed in the users' writing. Note there is no word-level highlighting feedback.

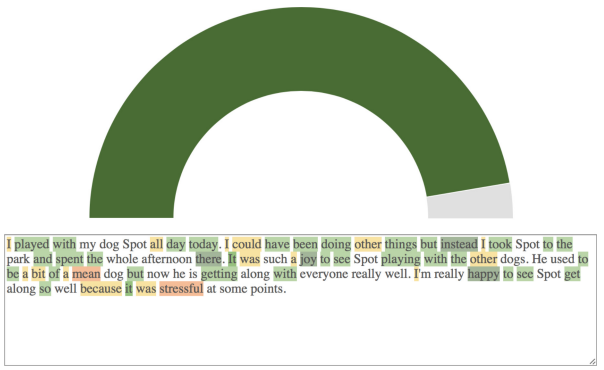

Fig. 2. E-meter Transparent Feedback: UI shows incremental word level feedback, with word color signaling emotional valence. Words that the system evaluates as positive are shown in green ("joy" and "happy") and words evaluated as negative are shown in red ("stressful" and "mean"). Meter bar position and color indicate that overall system evaluation of emotions expressed in the user's writing is highly positive.

working system to better engage and elicit direct user reactions to personally relevant data. Our system also provided real-time feedback that indicated to users that the system was authentic and not a deception [80]. But implementing a real-time transparent system imposes tradeoffs regarding performance of the underlying model. In particular, we needed to sustain performance across multiple hardware platforms. We therefore limited the complexity of our machine learning model and constrained our underlying feature set, so that it could run in real-time in Javascript within a browser.

3.1 Machine Learning Model

Little empirical work has examined how users respond to even simple operating models of transparency. Our current approach aims to gain direct feedback from users about an algorithm that interprets their personal data, and we have successfully deployed the approach in several prior studies [78, 80]. We chose a modeling approach that is accurate enough to be convincing but simple enough to explain to end-users. We therefore implemented a unigram based regression model. While such a model may be less objectively accurate than state-of-the-art methods using deep neural networks [21, 48, 58], it provides a straightforward operationalization of transparency that

is intuitively understandable while still appearing accurate and convincing to users [80]. And although the algorithm makes some errors, evaluating user reactions to these errors is an important aspect of our approach.

The emotion detection algorithm works as follows: Each word written by the user is evaluated for its positive/negative emotion association in our model. If that word is found in the model, then the overall mood rating in the system is updated. This constitutes an incremental linear regression that recalculates each time a word is written. In prior experiments [78, 80], users judged this algorithm to be accurate but not perfect; the median response for accuracy was "Accurate" ("6" on a 7-point Likert scale with the highest rating ("7") being "Very Accurate").

Models were trained on text gathered from the EmotiCal deployment [38, 79]. In that deployment, users reported on their activities and emotions over several weeks by writing textual entries about daily experiences and directly evaluating their mood in relation to those experiences. These data gave us a gold-standard supervised training set, where users labeled their own descriptions for underlying affect. We trained the linear regression on 6249 textual entries and mood scores from 164 EmotiCal users. Text features were stemmed using the Porter stemming algorithm [69] and then the top 600 unigrams were selected by F-score, i.e., we selected the 600 words that were most strongly predictive of user emotion ratings. Using a train/test split of 85/15 the linear regression tested at $R^2 = 0.25$; mean absolute error was .95 on the target variable (mood) scale of (−3,3). To implement this model on a larger range for the E-meter, we scaled the predictions to (0,100) to create a more continuous and variable experience for users. The mean absolute error of our model indicates that the E-meter will, on average, err by 15.83 points on a (0,100) scale for each user's mood prediction. As we noted above, users judged the algorithm to be accurate but not perfect.

This underlying model was evaluated by users for two different system versions: non-transparent and transparent.

Non-transparent version: As users wrote, the E-meter dynamically showed the system's overall interpretation of the emotional valence of their entire writing sample, in a visual meter (see Figure 1). If the overall text was interpreted negatively, then the gauge emptied to the left and turned more red. But if the text was judged to be positive, then the meter filled the gauge to the right and turned more green. Figure 1 shows a user writing sample that is judged to be very negative as indicated by the red on the far left of the meter. This continuous scale feedback represents the coarse global information that many machine learning systems currently display. Such systems give an overall rating but do not offer the user an insight into the detailed workings of the underlying algorithm.

Transparent version: In contrast, the transparent version provided fine-grained dynamic visual feedback revealing how the algorithm was making its overall prediction (See Figure 2). This transparency was operationalized as word-level feedback; in this system version, the E-meter signaled the affective weighting of each word the user types. Individual words are highlighted and color coded according to how the underlying algorithm interpreted that word's affect. If a word is associated with positive mood, then it will be highlighted dark green (e.g., "joy" and "happy"), whereas a word associated with negative mood will be highlighted red (e.g., "mean" and "stressful"). More neutral words were signaled as orange (e.g., "all" and "other"), and words that did not appear in our model were left transparent (e.g., "dog" and "Spot"). This incremental feedback allows users to see how each individual word they write contributes to the overall E-meter rating. Furthermore, words remain highlighted as users continue to type allowing users to assess each word's relative contribution to the overall score.

This form of transparency offers users insight into the underlying word-based regression model driving the E-meter evaluation; it depicts how the regression model correlates each word with positive or negative emotion to arrive at an overall weighting for the entire text that the user has

entered. The fact that the transparency visualization is persistent also allows users to scrutinize what they have written, reconciling the overall E-meter bar rating with the fine-grained word-level transparency.

Of course, we could have operationalized transparency in other ways. Other researchers have implemented transparency through natural language explanations [44, 85] and diagrams [46, 47, 61, 73]. However, in our case we can convey key aspects of underlying system mechanics through word highlighting. In addition, our operationalization allows the answering of counterfactual questions, an important part of explanation [57, 85]. Highlighting the text helps directly convey to the user what drives the algorithm and gives clear clues about the underlying linear model. In addition, by varying the colors of the highlighting we also show how the model is interpreting the specific words.

We now present four empirical studies of participants' reactions to such transparency. We explore users' preferences, evaluations, and perceived benefits for transparency information.

4 EXPLORING REACTIONS AND PREFERENCES FOR TRANSPARENCY

There has been relatively little exploration of users' *in situ* reactions to working algorithms that contain transparency features. Study 1 therefore compared transparent and non-transparent versions of an algorithm that analyzed the user's own personal data. For each system version, we assessed user accuracy judgments both before and after using the system. We also explored how these judgments related to user perceptions of trust and cognitive load, as well as their usage preferences.

4.1 Method

Prior to using E-meter, participants were shown an orienting video illustrating both transparent and non-transparent versions of the E-meter system and asked to predict system accuracy for each. Then they used both versions, generating a writing sample with each. After using each version, they were asked a series of follow-up questions, concerning perceived accuracy and trust. We also evaluated their perceived cognitive load. Users also provided their overall system preference after using both systems. The study design was counterbalanced; half the users experienced the non-transparent system version first. Questions and procedure were carefully piloted and had been used before in multiple prior studies. The study was approved by an Institutional Review Board.

4.1.1 Users. We recruited 100 users who had previously completed a subset of the Psychological General Well-being Index (PGWBI) [33], which was used as a screener. Users were recruited from Amazon Mechanical Turk and paid $3.33. The evaluation took 14.68 minutes on average. Following methodological recommendations concerning compliance of Mechanical Turk participants [18], we eliminated 26 respondents based on their responses to open ended questions, leaving us with a sample of 74 users.

4.1.2 Measures. Before actually using the system, participants saw short video animations of both transparent and non-transparent versions of the system. After viewing these they were asked to predict system accuracy for each. In each video, we simulated system behavior as filler Latin text was typed. We used Latin text, because we wanted high-level system comparisons that were not based on users' reactions to specific English words. Following each animation, we asked the *predicted accuracy* question: "This program evaluates the positivity/negativity of emotional experiences that users write about. How accurate or inaccurate do you think this program would be for you? The program works with English also." Users then provided qualitative explanations for their ratings*:* "Please give two or more reasons for the accuracy ratings you made on the previous page."

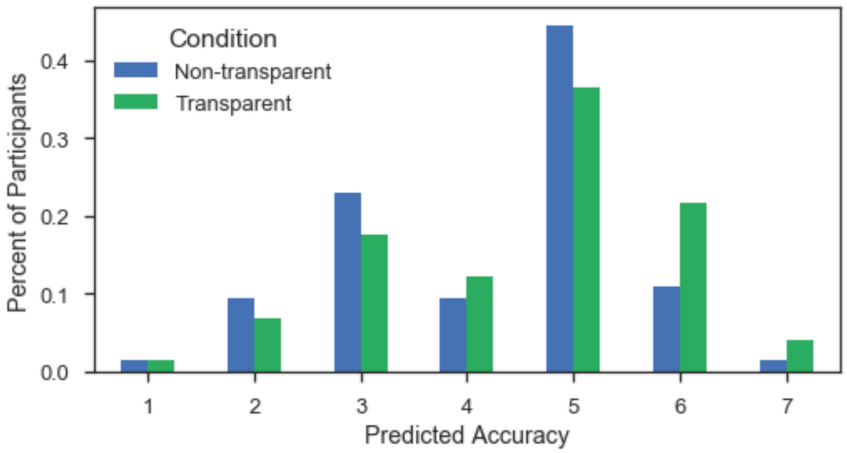

Fig. 3. Predicted Accuracy for Transparent versus Non-Transparent versions of E-meter in Study 1. Users predicted the Transparent version would be more accurate than the Non-Transparent version after viewing a video of it working but before actually using it.

Participants then began their writing activity. Following a protocol successfully deployed in prior work [76], users were presented with one of the two system versions with the instructions: "Please write at least 100 words about an emotional experience that affected you in the last week." After this system experience, users completed the TLX task load assessment [36]. Users then answered the questions to assess perceived accuracy and trust: "How positive or negative did you feel your writing was?" (*subjective affect*), "How positive or negative did the E-meter assess your writing to be?" (*system affect*), "How accurate or inaccurate was the E-meter in its assessment of your writing?" (*retrospective accuracy*), "How trustworthy or untrustworthy did you find the E-meter system?" (*subjective trust*). These questions were all 7-point Likert items. For example, the subjective trust questions items were from "Very Untrustworthy" to "Very Trustworthy." All questions had been successfully deployed in prior work [78, 80].

Users then repeated this process for the other system version. After using both versions, users answered a final *experience-based system preference* question "If you were to use the E-meter again, which system would you prefer?" They then supplied reasons for this: "Please give 2 or more reasons for the choice you made above."

4.2 Results

System perceived as moderately accurate and trustworthy: Overall, the median user found the E-meter to be "Slightly Accurate" and "Slightly Trustworthy," i.e., a rating of "5" on each 7-point scale. This *retrospective accuracy* evaluation is consistent with our previous work [80], where the median user found the E-meter to be "Accurate." There was no difference between conditions for either *retrospective accuracy* ($p = 0.24$) or *subjective trust* ($p = 0.41$).

Transparent version is predicted to be more accurate before usage: Before any hands-on experience with the system, participants generated *predicted accuracy* judgments for both the transparent and non-transparent system versions. Participants anticipated greater accuracy for the transparent system as indicated by Figure 3; for a paired *t*-test, $t(73) = 5.452$, $p = 0.022$, although the effect was small and means were 4.24 and 4.57, respectively.

Qualitative user comments support anticipated transparency benefits: On being asked to justify these predicted accuracy judgements, participants justified why they expected the transparent

system to be more accurate by drawing attention to the incremental color-coded feedback. They argued that this offered them a clearer sense of how the system worked, with this feedback boosting their confidence that the system was operating appropriately. Participant 50 (P50) wrote, "One meter is more transparent than the other. I can see how it works. I feel more confident in knowing exactly how it comes up with its answers. I tend to think it is more reliable." In the same vein, P73 wrote, "The [transparent system] had a legend with it and actually changed the color of the words I have written. … It was also more catchy and the colors stood out to me." Overall, then, before actually using either E-meter version, users anticipated that a system offering word-level transparency would be more accurate.

Retrospective accuracy predicts system preference: Recall that after experiencing each system version, we asked users for *retrospective accuracy*, and a final *experience-based system preference* question about which version they would choose for future usage. User perceptions of *retrospective accuracy* with both versions of the system highly correlate with their final *experience-based system preference* in a logistic regression model (p's = 0.019, 0.0001). Given only *retrospective accuracy* scores from both versions, we can predict the version choice with 69.5% accuracy in a fivefold cross-validated test. Therefore, knowing users' *predicted accuracy* was higher for the transparent version, we would expect that this would lead to users ultimately preferring the more transparent version of the system.

Transparency preferences disappear after usage: However, this positive predicted evaluation of transparent system accuracy did not persist after actual experience with the system when we analyzed final experience-based preferences.

We examined participants' written explanations of their experience-based preferences to better understand their choices. As we initially expected, many of those who preferred the transparent system after usage did so, because it illustrated the inner workings of the system. P72 said, "I know what the first [transparent] version is doing. I cannot tell what the second version [non-transparent] is doing. Because the second version does not give real feedback, I cannot make an informed decision when writing if I should be using it or not." P28 concurred, saying, "I think it [the transparent version] provides more engaging feedback and helps me better understand the reasons it gives for the amount in the meter."

However, to our surprise, the anticipated benefits of transparency did not generalize to all users. After experience with both systems, users were evenly split about which version of the system they preferred to use in the future: Fifty percent of participants (37) said that they would prefer the non-transparent version if they were to use the system again. The other 50% (37) chose the transparent condition. Consistent with these *experienced-based preference judgments*, participants also showed no overall differences in *subjective trust* after using the two different system versions ($t(73) = .910$, $p = 0.343$). Overall, both the final system choice and trust showed no differences between system versions, despite people being confident initially that the transparent version would be more accurate. What might explain these changed perceptions after usage?

Cognitive load did not explain preferences: One possible explanation for this changed perception is cognitive load. Transparency may demand attention and distract participants, in contrast to the non-transparent system version, which presents less information [12, 85]. Some participants' explanations for their experience-based preference seem to support this. These participants ($n = 13$) cited the distracting nature of the word highlighting. Users called word-based transparent feedback "annoying" (P3) and "obtrusive" (P7). P20 said that the non-transparent version was "a lot less distracting." P57 said "The individual highlighting of the words was distracting during writing; I wouldn't have minded it as much if I could turn it on and off." However, these subjective reports were not borne out by our quantitative analysis of cognitive load as assessed by the

TLX survey. A paired t-test comparing the overall TLX measures for both versions of the system indicated no difference in workload: $t(73) = -.05$, $p = 0.95$.

Reduced transparency may lead some users to overestimate system capabilities: Another potential reason why some users may ultimately prefer the non-transparent system relates to user inferences about algorithmic capability. Our qualitative analysis of participants' explanations for experience-based preference offers some insights into this. This analysis suggests that some users ascribe more advanced abilities to the non-transparent version. Nearly a quarter (24%) of users who chose the non-transparent E-meter as their preferred version stated that they preferred it because it took their overall writing context into account, incorporating information beyond simple lexical weightings. P66 said, "I think the [non-transparent version] takes into account everything you are writing and makes a decision better than just by focusing on word choice." P17 concurred, saying, "I like the second [non-transparent version] as it seems to focus on the whole and not each word." While such inferences about the non-transparent version are positive, they are also inaccurate. Recall that both systems use the same underlying machine learning model that uses solely individual word features.

What other reasons might explain this overestimation of system capabilities? Another possibility is that non-transparent feedback can hide low-level errors from users. In contrast, many transparent condition users identified highlighted words they felt were misclassified, leading them to downgrade their system evaluation. For example, P40 chose the non-transparent version, justifying their choice by saying, "...the biggest reason is that the most negative thought I had was expressed by the word 'isolated' in the text I wrote and the E-meter marked that one word as 'unimportant'—I couldn't get past that." P70 said: "Some associations don't make any sense, while others do." In contrast, non-transparent feedback did not expose these errors. If the algorithm was behaving consistently with their overall expectations, then users in the non-transparent condition judged it very positively.

4.3 Discussion

Our initial hypothesis was that providing detailed, transparent word-level feedback would be more helpful to users than the overall global predictions offered by non-transparent feedback. However, our first study unearthed some unexpected findings, showing that user interpretations of transparency feedback are far from straightforward. It is important to note that although the only objective differences between the systems lay in their transparency, users seemed to treat them as operating quite differently. Even though participants knew both versions of the system existed at the start of the experiment, they seemed to attribute very different qualities to each version after experiencing them.

Consistent with our initial expectations and the prior literature on transparency [28, 49, 85], users anticipated a preference for transparency before using the system. They expected that it would provide detailed incremental feedback about algorithmic decisions. But to our surprise, many participants did not retain this preference after using the system, at which point participants were evenly split between the systems in their trust and accuracy judgments. As we had originally anticipated, some users continued to prefer transparency; citing the increased insight that it offered into the algorithm's underlying operation. In contrast, others shifted their preference to the non-transparent version, but offered very different reasons for this judgment. While some users seemed to find transparency highlighting to be distracting, our cognitive load results do not support this. However, others preferred non-transparent feedback because it did not expose word-level errors, potentially leading users to overestimate the competence of the underlying algorithm with the consequence that they believed it to be more advanced than it was. For these users who preferred

the non-transparent version, it seemed that incremental feedback was providing more information than they required [12, 43].

Overall, users had better impressions of the transparent system initially, but those preferences disappeared after using both versions. Our observations suggest that the absence of detailed information in non-transparent feedback can hide errors; this error hiding leads some participants to form approximate, but positive, working heuristics about how the system operates. This may explain the lack of experience-based preference for transparent systems.

5 REACTIONS TO TRANSPARENCY AND EXPECTATION VIOLATION

Our first study suggests some unexpected and paradoxical effects for transparency. Although the transparent system was rated favorably before people used it, this preference was not maintained after they had direct experience with it. Transparency seemed to have mixed benefits: while feedback seemed to help some people form useful working models of system operation, others had their confidence in the system undermined by seeing system errors. At the same time, the non-transparent version may lead others to overestimate system capabilities, with a resulting preference for non-transparency.

We set out to explore these seemingly contradictory effects in a more targeted quantitative study. Again, we compared both forms of system feedback, but we also wanted to more directly explore potentially differing effects of transparency in relation to *expectation violation* [43]. While we anticipated overall benefits for transparency, prior work suggests system expectations play a critical role in system evaluations; if users feel the system is operating as they expect, then they judge it to be accurate and trustworthy [43, 80]. However, Study 1 indicated that these positive expectations might be undermined when transparency reveals errors. In Study 2 we therefore used expectation violation to assess users' beliefs about whether the system is behaving according to their expectations. Expectation violation was operationalized in this context as the difference between system judgments versus users' own evaluations of the emotional valence of their writing. Low violation means the system is behaving as expected, and high violation means there is a large discrepancy between user and system models.

Study 2 therefore explored relations between transparency, expectation violation and accuracy judgments. We expected that transparency would generally help but that reduced perceptions of accuracy could result from users observing system errors. We made two specific predictions:

- P1: Transparent interfaces should offer users insights into the underlying system, reducing expectation violation and increasing perceived system accuracy.
- P2: Transparency benefits depend on whether there is expectation violation. If users perceive the system is operating well, then transparency can undermine, reducing perceived accuracy.

5.1 Method

We used the same system and transparency manipulations as Study 1 with one difference. Study 2 is a between-participants design; so users only ever experience one form of system feedback, as we did not want prior system experience to influence users' judgments. Participants were randomly allocated to transparent (see Figure 2) and non-transparent (see Figure 1) conditions. As before, they were asked to write 100 words about an emotional experience in the E-meter system.

5.1.1 Users. We recruited 41 users from Amazon Turk and paid them $3.33. All had previously passed our standard psychological screener [30]. The evaluation took 13 minutes on average. This study was approved by an Institutional Review Board.

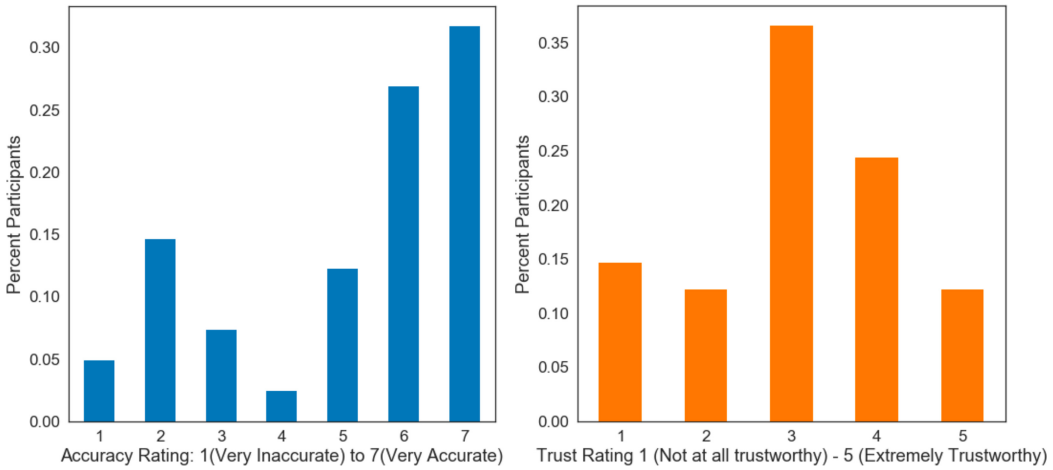

Fig. 4. Overall Accuracy and Trust ratings for E-meter in Study 2. Figure 4(a) (left) indicates that most users found the E-meter to be either "6," Accurate, or "7," Very Accurate. Figure 4(b) (right) indicates that most users found the E-meter to be either "3," Moderately Trustworthy, or "4," Very Trustworthy.

Table 1. Regression Showing Effects on Perceived Accuracy of Expectation Violation, Transparency, and Their Interaction for Study 2

	Coefficient	SE	Significance
Intercept	6.991	0.377	<0.0001
Expectation Violation	−1.736	0.601	<0.0001
Transparency	1.406	0.198	0.007
Transparency × Expectation Violation (Interaction)	−1.057	0.441	0.022

The overall regression is highly predictive $R^2 = .548$, $p < 0.0001$. Overall, Perceived Accuracy is low for High Expectation Violation. Transparency increases Perceived Accuracy, but the interaction term shows these benefits only occur when there is high Expectation Violation.

5.1.2 Measures. We asked the same questions as Study 1, with the addition of the following qualitative question: "Please explain how you think the system judges your writing."

5.2 Results

Participants complied with instructions to write at least 100 words, (mean words written = 107.74, sd = 14.8). Again, most users across conditions judged the E-meter to be "Accurate" or "Very Accurate" with the median being "Accurate" (a score of "6" on a 7 point scale) as shown in Figure 4(a). Users again found the E-meter to be "Moderately Trustworthy" ("4" on a 7 point scale) as shown in Figure 4(b).

We examined the effects of Transparency information using linear regression to model users' Accuracy judgments as dependent variable (see Table 1). We also included user's Expectation Violation as an independent variable. We also added an Interaction term (Expectation Violation × Transparency) in the regression model, because we predicted that the effects of Transparency would depend on Expectation Violation. Overall, then, independent measures in the regression model are Transparency, Expectation Violation, and the interaction between them. Expectation Violation was defined as the difference between the user's subjective affect judgment (their own assessment of the mood valence expressed in their writing) versus the system's overall assessment

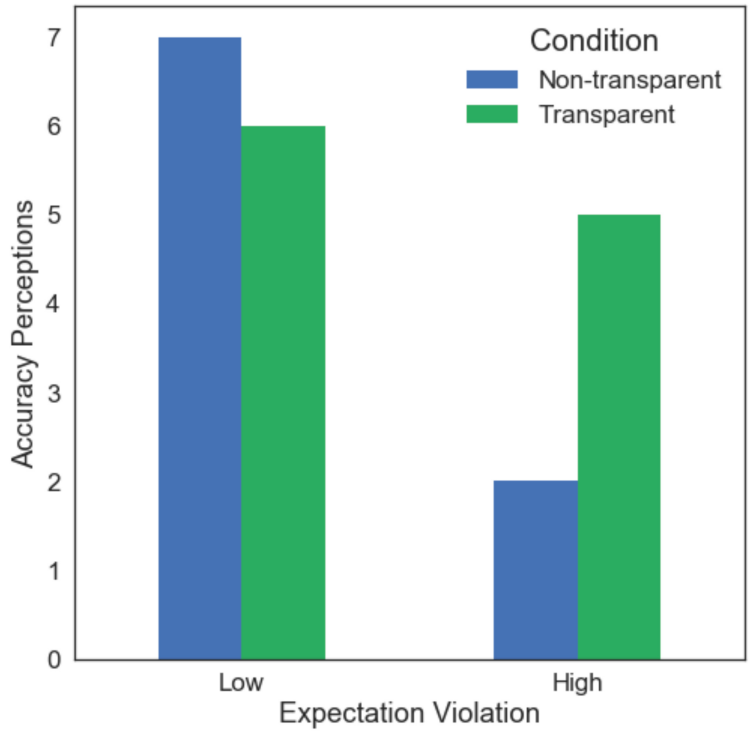

Fig. 5. Median Accuracy Perceptions under Low and High Expectation Violation for Study 2. Transparency and Expectation Violation interact to influence user Perceptions of Accuracy. When expectations are violated, Transparency helps to maintain positive perceptions of system Accuracy. When expectations are not violated, Transparency can reduce perceptions of Accuracy. Low Expectation Violation occurs when User and System emotion ratings differ by one or zero points on a 7-point scale. High Expectation Violation occurs when User and System emotion ratings differ by 2 or more scale points.

of the emotional valence that writing. A large difference indicates high Expectation Violation. The overall regression was highly predictive ($R^2 = .548$, $p < 0.0001$), and we now interpret its findings.

The regression model shows that Transparency information helps, but only with high Expectation Violation. As expected, both Expectation Violation and Transparency are associated with perceived accuracy. We see the predicted strong negative relation between Expectation Violation and Accuracy ($p < 0.0001$) confirming prior work [43, 80]. In other words, people rate the system as less accurate when it does not meet their expectations. In addition, as anticipated, Transparency has an overall positive effect on perceptions of accuracy.

However, this overall benefit of Transparency depends on Expectation Violation, as indicated by the interaction term in the regression. The interaction term valence shows that Transparency has the expected benefit when Expectation Violation is high (as shown in Figure 5). In other words, if people have high Expectation Violation, then Transparency increases perceived Accuracy. However, Transparency is not helpful for lower levels of Expectation Violation. Thus, Transparency only helps when people perceive the system not to be operating as they expected.

We next explored participants' qualitative responses to better understand Transparency effects, as well as the interaction between Transparency and Expectation Violation.

Transparency Helps Users Develop Mental Models of System Operation. As expected, many users exploited transparent feedback to form clear models of how the system operated. These

people were generally able to determine how the system weighted individual words to arrive at an interpretation, and this led them to view the system as accurate. In contrast, users in the non-transparent condition overall experienced more problems in understanding how the system operated. Many focused on major shifts in the movement of the E-meter in relation to their writing. However, this only allowed them to form an impressionistic global model of how the algorithm worked. This lack of clarity sometimes undermined their confidence in the system's accuracy. P4 felt that the system was inaccurate overall, but could not form a clear model about exactly why it was failing: "It was highly inaccurate because the experience was clearly a negative one, I specifically explained how awful I felt, I don't think that it could measure the sentiment of what I'm writing." Others also thought that the system was inaccurate, but the lack of transparency led them to propose erroneous accounts of how it was working. P11 said "it looked at length and speed of what I was typing"; P19 concurred, saying, "I thought it was only reacting to my WPM [words per minute]."

Overall then, users who saw transparency information seem better able to understand how the system is operating, which increases their accuracy judgments. Despite these overall benefits, however, some transparent system users seemed distracted or undermined by the system's interpretations of particular words. P28 downgraded their rating of the system's accuracy because of specific errors they had seen: "It went way down when I typed the word 'mad' but that was only a small part of the whole situation. The words that were good or bad seemed kind of arbitrary too." Similarly, P30 said: "I wrote, 'I was not thrilled' which is a negative statement, but this meter took the word 'thrilled' as a very positive thing." Despite overall perceptions that the algorithm was accurate, these examples suggest that for some users seeing word-level errors might paradoxically undermine previously positive expectations.

Transparency Can Undermine When Users Believe the System is Operating Well. To further assess these potential undermining effects, we examined cases of low expectation violation, i.e., instances of where users rated the E-meter as within 1 point of their own evaluation of their writing. By definition, the system is operating as expected for these users.

For some of these low expectation violation users, transparency seemed to create more questions than it answered, and the additional information provided by the word-level highlighting confused rather than clarified. One participant noted that while the system's final negative rating was consistent with their overall judgement of their own writing, the details of the highlighting did not make sense "...because the rating did not correspond to the number of identified words"; this user also noted "It gave a positivity rating of 1 [unanimously positive] even though it highlighted one or two words as red."

Other low expectation violation users seemed to experience a different type of problem on seeing transparency. For them the highlighting disconfirmed their system model, revealing that the system operated in an unexpected way. While non-transparent users were unable to discern this discrepancy, those who saw transparent feedback sometimes took issue with this. P41 said, "The key is to measure the overall emotional tone of the passage and it seems to fail at this." P36 said this simply: "I disliked that it cannot understand context." For these users, transparency revealed that the algorithm did not conform to their models of the task.

In contrast, some users in the non-transparent condition, seemed to retain high evaluations of the algorithm's accuracy, because they lacked such detailed information. The imprecise global feedback from the algorithm confirmed their overall interpretations, allowing them to build a consistent, if vague model of system operation. For example, P24 described how the system accuracy met their expectations: "I was writing about a negative topic and it continued to read in the negative state. The more upset I was writing, the further the dial went into the red." When

the non-transparent feedback matched their expectations, such users maintained high confidence that the system was accurate.

Transparency Helps When Expectations Are Violated. We also examined users who initially felt that the algorithm was violating their expectations, defined as where their evaluations differed from the system's by more than 1 point. For them, transparency had the anticipated benefit. It seemed to provide reassurance and explanation of the system behavior. One user, P27, noted that the system was relatively accurate, even if it violated their expectations: "I think that it was measuring words I used and rated them almost correctly." In the same vein, Participant 30 said "Even though it got several individual things wrong, I think it actually did a good job on the whole." For these users, transparency information can help moderate expectation violation, suggesting to users that the algorithm is operating in an interpretable manner, even if it does not exactly conform to their expectations.

5.3 Discussion

The second study provides new insight into transparency effects by indicating how transparency reactions are influenced by users' expectations about the system. We first confirm other work showing that as anticipated, transparency is beneficial overall [22, 49, 50]. However, we also show that transparency effects depend on user expectations [43]. Transparency has contradictory effects depending on the user's assessment of system performance. For those who were unsure about overall system accuracy, transparency boosts confidence in the system, by providing insight into how the algorithm operated. Paradoxically, this benefit did not accrue for users who felt that the system was behaving appropriately; transparency led them to notice word-level errors or see non-conformance to their mental models. Transparency may therefore have undermined their judgments of system accuracy.

6 REAL-TIME EFFECTS OF TRANSPARENCY

Study 2 elucidated the effects of transparency on people's judgments of algorithmic systems, showing how this depends on their expectations. However, it did not assess direct real-time effects of transparency on perceived accuracy, nor did it determine *when* users are most likely to benefit from transparency information. Study 3 examines the following two research questions:

- RQ1: When is transparency information most helpful? Is it most useful when users are developing models of system operation, or when they test well-defined accounts?
- RQ2: How does transparency change users' understanding of system operation?

6.1 Method

We again explored users' emotional writing behaviors using our E-meter platform, but the platform was further instrumented to gather real-time data. We wanted to directly measure the effects of transparency on users' system expectations and their perceived accuracy judgments. We therefore examine these effects dynamically at multiple time points, beginning when participants first used the system, but repeating these measurements at different points as users gained more system experience.

Our modified approach was the following. The E-meter was generally non-transparent. However, at three different points during their writing, users received a probe that prompted them to evaluate the accuracy of the (currently non-transparent) E-meter system. After responding to that accuracy probe, users saw transparency feedback showing how the algorithm had interpreted what they had written so far. They could review this information for as long as they wished. We timed how long participants reviewed this transparency information, and then probed how much

that information had improved their understanding of the underlying system. This procedure was repeated for two writing samples.

6.1.1 Users. We recruited 53 E-meter users who had previously passed a short screener (PG-WBI) [33]. Users were recruited from Amazon Turk and paid $3.33. The evaluation took 17 minutes on average. This study was approved by an Institutional Review Board.

6.1.2 Measures.

Real-time Expectation Violation: While writing, users received multiple prompts to elicit their judgments of the current emotional valence of their writing. On six occasions, users' writing was interrupted by a prompt asking them to make the following evaluation of their writing: "How positive or negative do you feel your writing is right now?" We then compared this user judgment with the current system emotion rating of their writing, and computed the discrepancy. We refer to this discrepancy as the *real-time expectation violation*, which is measured six times overall.

Transparency Viewing Time: After users had responded to each prompt, we showed them transparency information. They could view this for as long as they wished and we recorded this duration. We refer to this as the *transparency viewing time*. When users had finished viewing the transparent information they pressed a button labeled "Press this button to turn off highlighting and continue writing," which allowed them to continue.

Change in Understanding of the Algorithm: Before they recommenced writing, we asked users one further question. They were asked to evaluate their *understanding change* following their most recent viewing of the transparency. We used an ordinal 7-point Likert item outcome for understanding change from "Strongly Decreased My Understanding" to "Strongly Increased my Understanding."

We evaluated *real-time expectation violation, viewing time* and *understanding change* at three different points while the user was writing; when the user was 1/3rd through the writing task (i.e., they had written 30 words), when they were two-thirds complete (60 words), and then also when they were almost complete. Participants wrote about two separate personal experiences, so we enacted this procedure twice yielding six sets of judgments.

Overall Assessments: Finally, after using the system twice, participants were asked standard questions, as in Study 1 and 2. We probed their *overall perceptions* of the E-meter's accuracy, their *trust*, and their views of the system's operation.

6.2 Results

The study involves a within subjects design; we measure expectation violation 3 times per trial with 2 trials per experiment, so we analyze the data using a linear mixed model with crossed random effects from the R package lme4. The random effects control for intra-participant variance and intra-position variance (essentially order effects) and these are crossed, because each participant experiences each position. The fixed effect in the model is the Real-Time Expectation Violation, because we expected that users would spend more time examining transparency information when their expectations are violated. The dependent variable is the participants' Viewing Time for the transparent information. Before fitting the model, we take the natural log of the time variable, because it is distributed exponentially.

Real-Time Expectation Violation Predicts Viewing Time Spent Examining Transparent Information: The regression model is shown in Table 2. The model is highly predictive with $R^2 = .48$, $p < 0.0001$. The inclusion of the random effects in the model is significant at $p < 0.0001$, as determined by a likelihood-ratio test comparing the given model with a different model that does not include either the participant or position random effect. This comparison means there is significant variation between participants and also between the order effects within each participant.

Table 2. Regression Model Showing Expectation Violation Predicts Perceived Accuracy Judgments for Study 3

	Coefficient	Std. Error	Significance
(Intercept)	8.25	.16	<0.0001
Expectation Violation	0.07	.03	0.03

Model is highly predictive, $R^2 = 0.48$, $p < 0.0001$.

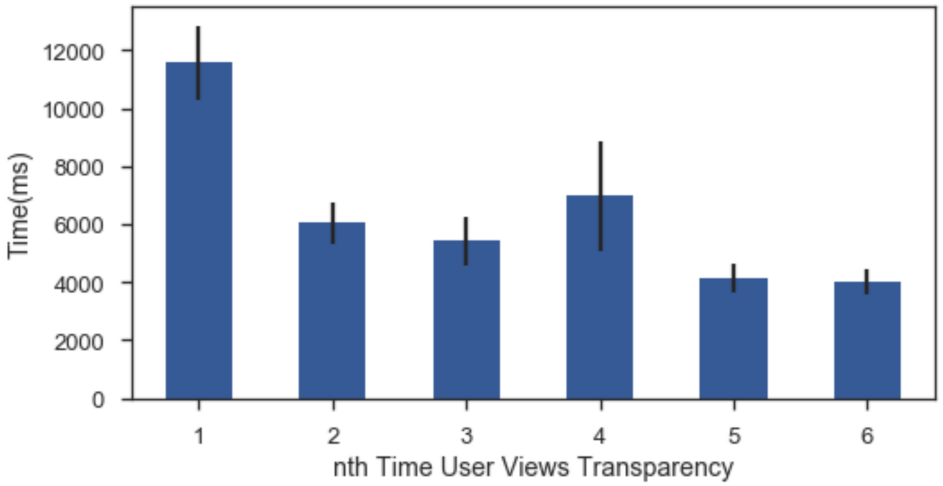

Fig. 6. Transparency Viewing Time in relation to Prompt Position for Study 3. People view Transparency information longer on the initial prompt. The increase in viewing time for Prompt 4 is likely due to the fact participants had just begun a second writing task (error bars show standard deviations).

As anticipated, Real Time Expectation Violation has a strong positive effect on Viewing Time. In other words, people view transparency information for much longer when their expectations are not met, although the effects are relatively small; a one unit increase in Expectation Violation is associated with a 7.25% increase in transparency Viewing Time (recall that time is measured on a log scale). Given that Expectation Violation ranges from [0,3], the maximum expectation violation is associated with a 23% increase in Transparency Viewing Time.

Transparency Viewing Time Decreases with Experience of the Algorithm: We also examined the effects of order on Transparency Viewing Time. As noted above, including the random effect for position was significant, indicating that there are significant differences in Transparency Viewing Time based on how long the participant had engaged with the algorithm. Our expectation was that Viewing Time would decrease as users gained more experience with the algorithm. Users have written just 30 words when they receive their first accuracy probe, and so we would expect them to have only a rudimentary model of how the algorithm operates, leading them to more thoroughly scrutinize the transparency information. We would also expect Viewing Time to decrease for subsequent prompts, as they develop a more refined system model.

Figure 6 plots Viewing Time according to prompt position. Results generally confirm our expectations. As anticipated, Viewing Time generally progressively decreases as people gain more experience with the algorithm. Most strikingly, the first position, where the user see transparency feedback for the first time greatly exceeds all other viewing times. Somewhat to our surprise, the fourth position seems to have a higher mean than the surrounding positions. However, this fourth

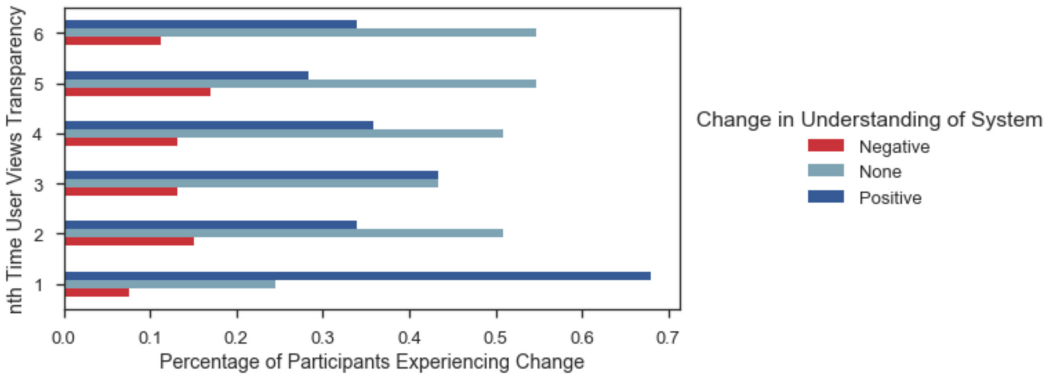

Fig. 7. Change in understanding of the algorithm in relation to prompt position for Study 3. Early viewing shows greater increases in understanding.

position corresponds to the first probe for the second writing task, and the elevation in Viewing Time may arise from the novelty of engaging in a second task with the algorithm.

Transparent Information Increases Understanding of the Algorithm. Recall that after each viewing, we asked users how viewing transparency information had changed their understanding of the algorithm (*understanding change*). Again we expected *understanding change* to depend on when people viewed the transparency information, with greater improvements for initial prompts when users have little prior exposure to the algorithm. By the same argument, we expected understanding change to decrease over prompts, as users built up a more informed model of system operation.

Results are shown in Figure 7. As expected, user responses to different prompts are significantly different using a Kruskal–Wallis test, $\chi^2(5) = 16.94$, $p < 0.01$. Visually examining the differences in Figure 7 shows that the first time users view the transparency, they generally experience positive changes in understanding. Subsequent viewings of transparent information are more neutral overall. On these later prompts, although a sizeable number of users increase their understanding, others feel their understanding is unchanged by the transparency. Confirming previous studies, a very small number of users find that their understanding decreases when viewing the transparency.

Overall the *understanding change* results mirror the *viewing time* data presented earlier. On early prompts, users are unclear about how the algorithm works, so that initial viewings promote greater understanding benefits. However both viewing time and understanding benefits progressively decrease for later prompts as users gain experience with the algorithm.

6.3 Discussion

We examined users' consumption of transparency information over time, as well as its real-time effects on their understanding of an algorithm. As expected, users spend progressively less time viewing transparency information as they gain experience with the algorithm. Furthermore, initial viewings promote distinct improvements in understanding, but these benefits decrease over time. Both viewing and understanding data suggest that consulting transparency information allows users to form clearer system models over time. Consistent with prior studies, we also find that users spend significantly more time viewing transparency information when their expectations are violated.

The previous three studies have explored the effects of transparency information, documenting how it interacts with expectation violation. They confirm that on some occasions transparency can be beneficial in reducing expectation violations, but on other occasions it can detract. However, these findings do not address questions concerning how we might better design transparent

systems. More specifically, they do not tell us exactly how to operationalize presentation of transparency information, given these findings. To explore potential design approaches, we need to better understand users' interactive experiences of transparency, which we explore in our final study.

7 DESIGNING FOR TRANSPARENCY

Ideally, transparency information should improve system understanding, without being undermining or distracting. However, it is difficult from our initial studies to know how to operationalize transparency in a practical way that achieves this. To better understand how to convey transparency to users in effective ways we employed a semi-structured interviewing process in our fourth study. We also used think-aloud methods to examine in depth how the type and timing of algorithmic transparency can inform decisions about how to design effective transparency. Studies 1 and 2 indicated that some users felt incremental feedback could sometimes provide too much information. Study 3 suggested that users' transparency needs differ according to their experience with the system. Study 4 therefore included a manipulation that examined users' reactions when they view increased transparency only "on demand," when they explicitly request it, and after they have finished writing. Overall, the study goals were to gather rich contextual qualitative data to further illuminate what factors influence the interpretation and uptake of transparency information, to motivate specific design approaches. We specifically asked participants how they were interpreting transparency feedback, which led several participants to suggest how that feedback might be improved.

7.1 Method

7.1.1 Users. Ten users were recruited from an internal participant pool at a large United States west coast university. They received course credit for participation. Participants' average age was 19.54 years (sd = 1.52) and 6 identified as female. This study was approved by an Institutional Review Board.

7.1.2 Measures. Users first completed a shortened version of the PGWBI [33], as screener. Users answered similar survey questions to Studies 1–3 but used in a think-aloud protocol; these questions were primarily used to prompt explanation and structure the interview. For this reason, we do not formally present the survey results here.

7.1.3 Procedure. The participants were randomly divided into one of two conditions. Both groups were given overall prediction feedback from the E-meter, but half had to deliberately request incremental visual feedback.

Passive Transparency: Five participants received real-time transparent word-level feedback about the algorithm's interpretation of their affect as they typed each word (as shown in Figure 2).

Requested Transparency: The other five saw non-transparent feedback (as shown in Figure 1), as they wrote. They only obtained transparent word-level feedback after they had finished the writing task; these users could explicitly request transparency feedback by clicking a button labeled "How is this rating calculated?"

The researcher explained the experiment and think-aloud procedure, demonstrating a think-aloud protocol on an email client. As in Studies 1–3, the researcher asked participants to write about a recent emotional experience. Unlike the prior studies, participants were asked to think-aloud while writing. Afterwards, the researcher conducted a semi-structured interview that included an on-screen survey. Passive transparency participants saw word-level feedback throughout. After the survey, participants in the requested transparency condition saw their writing again but with an added button labeled "How was this rating calculated?," which they pressed to reveal

the transparent word-level highlighting. They were asked to explain their reactions to the system and interpretations of transparency feedback. The entire process took around 50 minutes.

7.1.4 Analysis. Interviews were recorded using both audio and screen-recording. For this qualitative analysis, responses were coded using thematic analysis [11] specifically targeting how to support effective system transparency.

7.2 Results

Simplifying Transparency by Presenting Only High Impact Instances: Several users took issue with the level of detail provided by transparency, making spontaneous suggestions about how this might be modified. P10 felt that only the "big emotionally heavy words" should be highlighted. Other users felt similarly, P4 talked about select "trigger" words that that "trigger the foundation of the issue" and were essential to understanding the text. Likewise, P8 felt that there were a few important words, while the remainder just added noise: "it's taking into account words like 'stressful' and 'regretful' and stuff but then like everything else in between adds like an extra layer that complicates it." While the machine learning model was limited to the 600 most predictive words of mood, that model still seemed to present too many extraneous words that these users felt were unimportant. It may be that users need to identify a small number of clear examples of words showing strongly positive and negative affect, to form a working model of the system's operation.

Transparency May Violate User Expectations Even When the System Is Correct: Just as some participants felt there were many extraneous highlighted words, other participants focused on specific words that they judged had been misinterpreted by the system. Users wrote about their emotional experiences in the first person, often using the word "I." In our machine learning model, "I" has a slightly negative connotation that confused our users, because many of them thought "I" should be neutral. P8 said "So I was gonna say that yellow words would be neutral because it has highlighted 'I'...." Along the same lines, when analyzing the highlighting of different words, P6 said "'I'? Mmm, I don't understand that either." P5 even started conjecturing about the actual system model saying, "'I' doesn't seem like it would have... [participant trails off] unless people speak in objective terms when they're talking about more positive experiences."

User Heuristics May Contradict Transparency: Other problems arose, because users formed working heuristics of how the system operated, which were sometimes contradicted by word-level feedback. Four users felt that there were discrepancies between the overall rating and the transparent word-level feedback within the system. When P8 viewed the word-level transparency they started off by saying, "I'm very confused" and then explained how they felt the overall rating should just be calculated as a ratio of positive/negative words—"Well I just assumed that some words would be coded as positive or negative and then it would just like do a ratio of those two." P6 explained it similarly, the overall rating showed a slightly negative emotion rating for their written passage but, in P6's words: "So when you look at the comparison with the meter, and you look at my paragraph itself, right? There's more words that are highlighted in green."

These participants seem to be using simple heuristics to relate the overall document rating and word-level transparency. They feel that the overall rating should reflect a simple ratio of the number of positive versus negative words in the word-level transparency. For example, if the highlighted words are primarily green, then the overall rating should be very positive. However, in our machine learning model, a single word such as "angry" could be rated negatively enough that it would cancel out multiple mildly positive words. Some users arrive at this correct model after consciously engaging with the system. For example, recall how P6 talked about how they felt the word-level highlighting and the overall rating were incongruent; however, after thinking more deeply, this participant later said: "... it's weighing certain words, right? Because obviously these

two words, right? Like "upset" and like—er yeah, "upset" really polarize the meter." This statement indicates that this user has moved beyond their initial heuristic that all words are weighted equally; they are now noting how one word has a larger effect in the system.

7.3 Discussion

Study 4 again reveals the complexity involved in presenting transparency information. Confirming our first three studies, we again showed that providing more detailed transparency information isn't always better. Together, these observations again suggest that users initially form simple working hypotheses about system operation. Users seem to engage with transparency first by operating with these simple hypotheses and only critically re-examine these when their expectations aren't met. As with other areas of reasoning, it may be that when interacting with a system, users first engage in rapid, approximate, System 1 thinking and only engage in deeper, more analytic, System 2 thinking when directly prompted or confronted with clearly anomalous information [41, 67].

While some users saw potential benefits to transparency information, others argued against "complete" transparency, preferring to see only a subset of "important" words. This observation suggests a principle for transparency presentation that restricts the amount of information presented. This might involve explaining as much variation as possible using the smallest number of explanatory features. We might therefore aim to weigh the overall number of features we present against the information they provide. This is consistent with machine learning approaches to developing models with high dimensional feature sets that aim to identify features with the greatest explanatory power. While we know of technical methods that support this [56], we have not seen such approaches to transparency actually tested with users.

A second observation is that users may take issue with detailed transparency and predictions, even when underlying system models are objectively correct. We saw this with interpretations of the word "I," which some users felt should not have negative connotations. However, extant literature confirms that system feedback is correct, as high usage of first-person singular pronouns is correlated with depression and negative mood [66, 74]. Such examples indicate a problem when system feedback reveals information that contradicts the users' expectations. Even if the system is objectively correct when displaying transparency information, it can still cause users to take issue and result in poorer perceptions of the system.

This creates quite a difficult problem for system designers. If it were possible to know which features users prompt mistaken beliefs, then these features could be excluded from transparency. Unfortunately, short of testing user beliefs about all features, this may be very hard to do. We also saw that users don't necessarily interrogate transparency information to deeply analyze all its implications. Instead, users often look for quick heuristic routes to confirm or discredit simple working theories. We should therefore design intelligent systems in ways that are consistent with social science findings [41, 67], by allowing users to develop simple working heuristics but also invite them to evolve more accurate mental models when they are motivated to do so.

8 OVERALL DISCUSSION

Unlike much technically oriented work that aims to develop new transparency algorithms, this article explores user reactions to working transparent algorithms. Specifically, we examined users engaging with a real system that interprets self-generated personally relevant emotional data. We deployed a necessarily simple algorithm that users find accurate [78, 80], to create simple transparent visual feedback that conveys the underlying operation of the algorithm. All four studies indicate that developing and deploying transparent smart systems is complex in practice. We observed unexpected user reactions to our attempts to provide detailed information

about algorithmic operation. In Study 1 we found that before actual usage, participants initially anticipate greater accuracy for the transparent version of the E-meter, but this is altered by their system experiences. After using both system versions, and although both versions of the algorithm were perceived as accurate, users were split 50–50 in their preference. Trust data showed similar ambivalence. We suggested several possible reasons for this shift that were confirmed in Study 2. Study 2 showed that transparency information was generally helpful, but that its specific effects depended on expectation violation. Participants experiencing high expectation violation (i.e., those who were undergoing difficulties forming a model of system operation), derived benefits from transparency information. Feedback allowed them to see how system judgments were based on word-based weightings. In contrast those with low expectation violation, who had solid models of system operation sometimes had their judgments undermined by transparency information. Their confidence in the system was lowered by seeing system errors, or when the system model did not conform to their own model.

Study 3 confirmed these prior results and deepened our understanding of transparency, again showing that its benefits depend on context. It indicated that users are most receptive to such explanatory information as they form models of system operation during their early experiences with the system. Furthermore, transparency induces greater understanding benefits during initial phases of system usage.

Although many user comments in Study 1 mentioned how incremental word level affective feedback was highly distracting, these comments were not reflected by a reliable overall difference in cognitive load between transparency conditions as assessed by NASA TLX. However more sensitive real-time measures of cognitive load might yield different results.

These results raise the question of how we can operationalize transparency in ways that don't distract, while simultaneously allowing users to engage with the system using simple heuristics and facilitating advanced understanding. Study 4 gathered richer contextual data to illuminate exactly how to operationalize transparency to fit the goals of Studies 1–3. To improve clarity, we discovered that some users prefer to see only major contributing features to the overall algorithm rating. More detailed transparency information can lead some users to falsely believe the system is operating incorrectly. Furthermore, users often evaluate transparency using simple heuristics rather than deep reflection. Together these results suggest that supporting transparency is complex and there are myriad decisions that affect the user experience when deciding how to operationalize it. We now discuss future design approaches that build on these observations, and then move onto theoretical implications.

8.1 Meeting the Competing Needs of Transparency Through Progressive Disclosure

Our studies reveal requirements that transparency must meet to be effective for users. This is a challenge, because some of these requirements seem internally inconsistent. How can we allow users to develop preliminary working heuristics, while later facilitating detailed understanding for those users who value it? One design solution is suggested by an interaction that took place in Study 4. Recall that in Study 4, some users wrote about their emotional experience using the non-transparent version and only later saw the word-level transparency after clicking a button labeled "How was this rating generated?" After clicking the button to reveal transparency, P11 had further questions. Quite naturally, the participant pointed at the button again and had the following exchange with the researcher:

P11: Can I click this? Does this...?

Interviewer: I don't think it shows any more than that.

P11: Damn it.

Interviewer: Yeah. If you were to click it, what would you expect to see more about?

P11: I want bullet points to tell me why it works the way it does.

Even after seeing the exhaustive transparent feature contributions to the overall rating, this user still had outstanding questions about the system's inner workings. This datapoint suggests an interaction paradigm that meets the competing needs these four studies have generated: *progressive disclosure*.

Progressive disclosure has a long history in UI research, dating back to the Xerox Star and early word processing systems [15, 62, 76]. The original concept involved hiding advanced interface controls; allowing users to make fewer initial errors and learn the system more effectively [15]. In other words, advanced information and explanation is provided on an "as needed" basis, only when the user requests it.

We can apply the principles of progressive disclosure directly to transparency in intelligent systems. For example, similar to Study 4, the E-meter could show a "How was this rating calculated?" button. In this setting, the E-meter might start with a document-level rating, which reduces distraction and avoids unnecessary complexity. Upon first press of the button, the E-meter might show a brief natural language explanation, e.g., "This rating was calculated using the positive and negative weighting of the words you have written." Assuming that this did not satisfy the user's explanatory needs, on a subsequent press the E-meter might show a subset of the high confidence and high impact words. This second press would satisfy the users in Study 4 who only wanted to see the major factors influencing the algorithm. However, some users (like those in Studies 2 and 3) may want yet more transparency. Yet another press of the button could reveal further features that contribute to the overall score. Further presses could provide details about how the training data was collected, or a textual summary of the machine learning model.

In this progressive disclosure approach, following UX and human conversational principles, explanation is presented as two-way communication with the user determining exactly when and how explanations are provided [29, 32, 75]. Note, too, that because transparency is provided "on demand" this removes confusions and inefficiencies arising from spurious, unwanted explanations, and adjusts explanations to users' requirements.

Taken together, our findings suggest important potential implications for the design of transparent systems that exhibit progressive disclosure. To comply with progressive disclosure design principles, systems must offer:

- *"On-demand" information provision.* Users should drive the overall interaction. Users must therefore be able to control both when transparency information is provided, as well as the complexity of the information provided. This is a considerable design challenge. The "on demand" requirement also demands novel presentational approaches. Transparency information needs to both be readily accessible, while at the same time remaining unobtrusive, if users' explanatory needs are being met.
- *Hierarchically organized explanatory information.* Progressive disclosure requires a hierarchy of explanatory information that ranges from simple to complex. More work is required to better understand and define these different levels of transparency in relation to a given application domain. It is intuitive that explanations involving linear relations among small numbers of features will be easier to understand than those characterizing complex weightings of large-scale engineered feature sets. Beyond this, however, more theoretical work is needed to systematically define such levels, and validate them against actual user understanding.
- *Context tracking.* The system needs to keep track of the context and the information that has already been provided to the user, to better meet the users' current informational needs,

following conversational principles. Initial explanations should be relatively straightforward, but the system must track what explanatory information the user has already been exposed to.

We believe that such Progressive Disclosure of transparency information is broadly applicable to other types of intelligent system, and we now describe how these principles could be more generally applied. For example, other researchers have examined transparency in the context of models predicting patient outcomes in the medical domain [1, 46]. In this context, patients have a risk of a disease that is determined by a large number of predictive features (e.g., previous medical diagnoses, number of doctor visits, type of care, and so forth). Each feature could be visualized and ranked for its contribution to the overall predicted diagnosis risk score. In addition, users could compare outcomes between different patients and conduct what-if analyses by modifying patient attributes and seeing how predicted risk changes. However, our results suggest that exposing such feature complexity immediately could overwhelm certain users, leading them to reject the diagnostic tool. To operationalize progressive disclosure in this setting, we might present the predicted risk score by default using a short natural language explanation. Should the user request more information, we can incorporate visualizations of features that most contribute to the predicted risk score [1, 73]. A request for further information could show similar patients and previous outcomes that have informed the current prediction. Finally, if the user continues requesting increasing amounts of disclosure about a prediction, we can infer that user is invested enough to truly engage with the system and we can present a model that can be explored (as in the explorable predictions of Hollis et al. [38]). This hypothetical reasoning tool might allow users to modify the current patient's symptoms and see how this changes the risk prediction, helping the user to consciously build an accurate mental model of the machine learning predictions. Again, the major contribution of progressive disclosure is avoiding overwhelming users with information, but instead slowly increasing transparency, and exposing more of the underlying model as users indicate a willingness to engage meaningfully with it. We believe this can both support transparency and improve the acceptance of intelligent systems in many realms.

One potential objection to Progressive Disclosure is that exposing users to initially simple models may inhibit learning of complex, but more accurate models. However, empirical studies examining progressive disclosure show that people learn better when they are initially limited to a smaller set of core features. This enhanced understanding of core concepts gives them a framework that later helps them integrate the complexities of the advanced features [6, 15].

8.2 Transparency Theory and Methods

In addition to these design implications, we also contribute to theoretical framings of transparency. Our work suggests that users' transparency requirements are not monolithic but highly context-dependent. Transparency needs evolve within each user, as that user refines their model of how the algorithm is operating. Consistent with other research [43, 50, 51], our results show that providing overdetailed information can overload or undermine a user who is operating with a simple model.

Our work also underscores the importance of expectations in determining reactions to transparency. On the one hand, transparency benefits those with high expectation violation who are still developing a model of system operation. On the other hand, the same information may erode the confidence of users who have already developed a working model of system operation. These findings confirm other work showing relations between system perceptions and expectation violation [43, 80].

In more general terms, these results link to social science theory that characterizes whether and when to provide transparency information. We found that transparency is helpful when people

don't understand, but it can undermine when expectations are met. This is also consistent with literature on explanation from the social sciences, which argues that in human–human interaction, explanations are "occasioned," being provided only when the situation demands it [29, 37, 75]. This overlaps with philosophical framings suggesting that explanations involve counterfactuals, i.e., a comparison between an unexpected but actual situation, against an expected but non-actual one [37, 57]. Likewise, our recommendations about progressive disclosure coincide with theory that argues that conversational misunderstandings are repaired on an incremental basis, with the exact explanation depending on the current understanding of interactional participants [75].

We found that some users relied on simple heuristics and did not welcome complex explanations unless the situation seemed strikingly anomalous [41, 43, 67]. These findings are consistent with recent work on folk theories in algorithmic systems [20, 24] as well as theoretical work on decision making. For example, Kahneman's dual process decision theory as well as elaboration likelihood models (ELM) of persuasion both show that people commonly operate with simple (often approximate) situational heuristics [41, 67]. ELM also characterizes these approximate heuristics for evaluating persuasive messages. Heuristics involve focusing on global message attributes (e.g., the identity of the speaker, or the length of the message) rather than a detailed examination of how the message content relates to prior knowledge or beliefs [67]. Likewise, we observed that some users adopted approximate models of how the algorithm operated, e.g., believing that emotion ratings were determined by the relative proportions of positive versus negative words, or by the presence of specific emotionally laden words. These overlaps suggest that persuasion and dual process theories could offer important insights for designing transparent algorithms. These theories characterize both when users rely on simple heuristics versus more complex cognitive models, as well as the nature of the underlying heuristics that users adopt.

Another complex theoretical issue arising from our results concerns error. Future theory needs to focus on the effects of different types and levels of errors. Our findings suggest that transparency may have paradoxical effects; exposing users to system errors may undermine confidence in users who believe the algorithm is operating well [60]. Our studies did not systematically vary error, however, and prior work indicates that reactions to system explanations depend on the level of error [50, 51]. A second complex issue concerns differences between objective versus perceived error. In our studies we saw instances where users falsely believed that the system was inaccurate. For example, they challenged system interpretations of words such as "I," which are objectively associated with negative affect [66], but our users believed should be neutral. We also saw the opposite pattern of failing to detect actual errors. On other occasions, such as in non-transparent settings, users sometimes overattributed competence to the system. Here some falsely believed that the system was operating in a more sophisticated, context-sensitive manner than it actually was. It is critical that we better understand these phenomena, as there are significant implications. Other work has shown that users will modify high stakes personal health behaviors based on false inferences about an algorithm's high accuracy [39, 70].

Our findings also point to the importance of individual differences. It was apparent that different users have varying reactions to, and expectations about intelligent systems. One possibility is that these differences arise from individual user traits, such as need for control [2] or need for cognition [14]. Future work should examine the relation between such individual traits and reactions to transparency, allowing designers to profile users and deploying personalized system versions.

Our results also have important methodological implications. One method used in prior studies is to provide potential system users with hypothetical scenarios describing system operation and eliciting reactions to those systems. These methods offer ways to collect controlled user data at scale [28, 50]. However, results from Study 1 indicate the importance of direct user experience

when making system evaluations; users' perceptions of the system were very different following actual usage compared with their projected reactions prior to usage. Care needs to be taken with usage of scenario-based methods.

8.3 Limitations

The current research examines the important algorithmic domain of emotion analytics, but clearly other contexts need to be explored. Furthermore, our deployment of a working algorithm meant that results were obtained for situations where our algorithm generated moderate numbers of errors—future research should compare contexts where there are different levels of errors [50, 51]. Additionally, while users generated their own data in our system, results were not directly used to inform other aspects of the user's personal behavior, as in Hollis et al. [38], so the costs of system errors were low. While this is appropriate for exploring understanding of initial algorithms with moderate error rates, future work might explore user reactions to transparency in higher risk contexts. Our work also derived these insights from one operationalization of transparency, namely a dynamic visualization of an algorithm. There are many other ways to illustrate how an algorithm operates including verbal explanations, concrete user exploration, and so forth [44, 45, 51, 73, 88].

These are clear limitations of the work that naturally raise significant questions about generalizability of our findings to other contexts. Future work therefore needs to examine more complex machine learning models, different designs and different user expertise. Nevertheless, we believe that important aspects of our results are likely to generalize. First, our results replicate and extend other studies showing that algorithm accuracy evaluations closely relate to expectation violation [43, 78, 80]. We also confirm other work showing that revealing algorithmic errors can undermine users' perceptions of accuracy [50, 51, 60, 81]. Some of these other results were obtained with different learning models in settings that differed from the current study, and in contexts where users were not always domain experts. And although we explored extremely simple learning models in this study, our results may be relevant to more complex modeling methods. Newer technical approaches attempt to increase intelligibility of complex neural models by approximating them using simpler more intuitive models [47, 53, 54]. So, while our results do not apply directly to neural methods, they may become more relevant, as such approximation methods become more common. Having said this, further empirical work is clearly needed to directly explore users' sensemaking and understanding of such complex neural models, especially when they are not domain experts. Future work also needs to tackle the considerable design challenges involved in creating progressive disclosure interfaces that provide timely useful explanations that do not overwhelm or undermine the user.

9 CONCLUSION

Overall, our data suggest empirically motivated challenges in designing effective UX methods to support transparency for complex algorithms. Our results also reveal potential new research questions regarding user traits, system heuristics, and errors. The current study indicates a promising design approach involving progressive disclosure that we intend to explore in future work. It is critical to answer these questions as we continue to deploy intelligent systems with increasing ubiquity and impact.

ACKNOWLEDGMENTS

We thank our research assistant, Jack Bauman, for his valuable help with this research. We also thank our research participants for sharing their time and thoughts with us.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. 1–18. https://doi.org/10.1145/3173574.3174156

[2] Icek Ajzen. 1991. The theory of planned behavior. *Organiz. Behav. Hum. Decis. Process.* 50, 2 (1991), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

[3] Julia Angwin and Jeff Larson. 2016. Machine bias. *ProPublica*. Retrieved October 27, 2017 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[4] Stavros Antifakos, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. 2005. Towards improving trust in context-aware systems by displaying system confidence. In *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services (MobileHCI'05)*. 9. https://doi.org/10.1145/1085777.1085780

[5] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. "What is relevant in a text document?": An interpretable machine learning approach. *PLoS One* 12, 8 (2017), e0181142. https://doi.org/10.1371/journal.pone.0181142

[6] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and accountability: Human considerations in context-aware systems. *Hum.-Comput. Interact.* 16, 2 (2001), 193–212. https://doi.org/10.1207/S15327051HCI16234_05

[7] Frank Bentley, Konrad Tollmar, Peter Stephenson, Laura Levy, Brian Jones, Scott Robertson, Ed Price, Richard Catrambone, and Jeff Wilson. 2013. Health mashups: Presenting statistical patterns between wellbeing data and context in natural language to promote behavior change. *ACM Trans. Comput.-Hum. Interact.* 20, 5 (2013), 1–27. https://doi.org/10.1145/2503823

[8] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "It's reducing a human being to a percentage"; Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. 1–14. https://doi.org/10.1145/3173574.3173951

[9] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*. 4349–4357.

[10] Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics Inf. Technol.* 15, 3 (2013), 209–227. https://doi.org/10.1007/s10676-013-9321-6

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[12] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important?: A study of deployed, low-cost intelligent interactive systems. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*. 169–178.

[13] Moira Burke, Brian Amento, and Philip Isenhour. 2006. Error correction of voicemail transcripts in SCANMail. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'06)*. ACM, New York, NY, 339–348. DOI: http://dx.doi.org/10.1145/1124772.1124823

[14] John T. Cacioppo, Richard E. Petty, and Chuan Feng Kao. 1984. The efficient assessment of need for cognition. *J. Pers. Assess.* 48, 3 (1984), 306–307. https://doi.org/10.1207/s15327752jpa4803_13

[15] John M. Carroll and Caroline Carrithers. 1984. Training wheels in a user interface. *Commun. ACM* 27, 8 (1984), 800–806. https://doi.org/10.1145/358198.358218

[16] CCS Insight. 2016. CCS Insight Wearables End-user Survey. Retrieved August 11, 2018 from http://www.ccsinsight.com/.

[17] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding self-reflection: How people reflect on personal data through visual data exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth'17)*. 173–182. https://doi.org/10.1145/3154862.3154881

[18] Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems (CHI'14)*. 1143–1152. https://doi.org/10.1145/2556288.2557372

[19] Mary L. Cummings. 2004. Automation bias in intelligent time critical decision support systems. In *Proceedings of the AIAA 1st Intelligent Systems Technical Conference*. 557–562.

[20] Michael A. DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. 2018. How people form folk theories of social media feeds and what it means for how we study self-presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. 1–12. https://doi.org/10.1145/3173574.3173694

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs]. Retrieved December 27, 2018 from http://arxiv.org/abs/1810.04805.

[22] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *Int. J. Hum.-Comput. Stud* 58, 6 (2003), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

[23] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI'18)*, 211–223. https://doi.org/10.1145/3172944.3172961

[24] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First i like it, then i hide it: Folk theories of social feeds. In *Proceedings of the 2016 cHI Conference on Human Factors in Computing Systems*. 2371–2382.

[25] Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. Communicating algorithmic process in online behavioral advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. 432:1–432:13. https://doi.org/10.1145/3173574.3174006

[26] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that i wasn't really that close to [Her]": Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*. 153–162. https://doi.org/10.1145/2702123.2702556

[27] Mads Frost, Afsaneh Doryab, Maria Faurholt-Jepsen, Lars Vedel Kessing, and Jakob E. Bardram. 2013. Supporting disease insight through data analysis: Refinements of the monarca self-assessment system. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, New York, 133–142. https://doi.org/10.1145/2493432.2493507

[28] Pedro García Garcia, Enrico Costanza, Jhim Verame, Diana Nowacka, and Sarvapali D. Ramchurn. (in press). Seeing (Movement) is believing: The effect of motion on perception of automatic systems performance. To appear in *Hum.–Comput Interact*. https://doi.org/10.1080/07370024.2018.1453815

[29] Harold Garfinkel. 1991. *Studies in Ethnomethodology*. Wiley.

[30] K. Goddard, A. Roudsari, and J. C. Wyatt. 2012. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inf. Assoc.* 19, 1 (2012), 121–127. DOI: 10.1136/amiajnl-2011-000089. PMC 3240751. PMID 21685142.

[31] Bryce Goodman and Seth Flaxman. 2017. European union regulations on algorithmic decision-making and a "right to explanation." *AI Mag.* 38, 3 (2017), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

[32] H. P. Grice. 1975. Logic and conversation. In *Syntax and Semantics*, P. Cole and J. Morgan (Eds.), Vol. 3. Academic Press.

[33] E. Grossi, N. Groth, P. Mosconi, R. Cerutti, F. Pace, A. Compare, and G. Apolone. 2006. Development and validation of the short version of the Psychological General Well-Being Index (PGWB-S). *Health Qual. Life Outcomes* 4, 88 (2006). DOI: 10.1186/1477-7525-4-88

[34] Chloe Gui and Victoria Chan. 2017. Machine learning in medicine. *Univ. West. Ont. Med. J.* 86, 2 (2017), 76–78. https://doi.org/10.5206/uwomj.v86i2.2060

[35] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. arXiv:1610.02413 [cs]. Retrieved from http://arxiv.org/abs/1610.02413.

[36] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (task load index): results of empirical and theoretical research. In *Advances in Psychology*, Peter A. Hancock and Najmedin Meshkati (eds.). North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[37] Denis J. Hilton. 1990. Conversational processes and causal explanation. *Psychol. Bull.* 107, 1 (1990), 65–81. https://doi.org/10.1037/0033-2909.107.1.65

[38] Victoria Hollis, Artie Konrad, Aaron Springer, Chris Antoun, Matthew Antoun, Rob Martin, and Steve Whittaker. 2017. What does all this data mean for my future mood? actionable analytics and targeted reflection for emotional well-being. *Hum.–Comput. Interact.* 32, 5–6 (2017), 208–267. https://doi.org/10.1080/07370024.2016.1277724

[39] Victoria Hollis, Alon Pekurovsky, Eunika Wu, and Steve Whittaker. 2018. On being told how we feel: How Algorithmic sensor feedback influences emotion perception. *Proc. ACM Interact. Mob. Wear. Ubiq. Technol.* 2, 3 (2018), 114:1–114:31. https://doi.org/10.1145/3264924

[40] Ellen Isaacs, Artie Konrad, Alan Walendowski, Thomas Lennig, Victoria Hollis, and Steve Whittaker. 2013. Echoes from the past: How technology mediated reflection improves well-being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, 1071–1080. https://doi.org/10.1145/2470654.2466137

[41] Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Macmillan.

[42] SeungJun Kim, Jaemin Chun, and Anind K. Dey. 2015. Sensors know when to interrupt you in the car: Detecting driver interruptibility through monitoring of peripheral interactions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*. 487–496. https://doi.org/10.1145/2702123.2702409

[43] René F. Kizilcec. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, 2390–2395. https://doi.org/10.1145/2858036.2858402

[44] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI'15)*. 126–137. https://doi.org/10.1145/2678025.2701399

[45] Todd Kulesza, Weng-Keen Wong, Simone Stumpf, Stephen Perona, Rachel White, Margaret M. Burnett, Ian Oberst, and Andrew J. Ko. 2008. Fixing the program my computer learned: Barriers for end users, challenges for the machine. In *Proceedings of the 13th International Conference on Intelligent User Interfaces (IUI'09)*. 187. https://doi.org/10.1145/1502650.1502678

[46] Bai. C. Kwon, M. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun, and J. Choo. 2019. RetainVis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE Trans. Vis. Comput. Graph.* 25, 1 (2019), 299–309. https://doi.org/10.1109/TVCG.2018.2865027

[47] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & explorable approximations of black box models. arXiv:1707.01154 [cs]. Retrieved 18, 2018 from http://arxiv.org/abs/1707.01154.

[48] Shoushan Li, Lei Huang, Rong Wang, and Guodong Zhou. 2015. Sentence-level emotion classification with label and context dependence. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1045–1053.

[49] Brian Y. Lim and Anind K. Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th International Conference on Ubiquitous Computing*. 195–204.

[50] Brian Y. Lim and Anind K. Dey. 2011. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th International Conference on Ubiquitous Computing*. 415–424.

[51] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI'09)*. 2119. https://doi.org/10.1145/1518701.1519023

[52] Zachary C. Lipton. 2016. The mythos of model interpretability. arXiv:1606.03490 [cs, stat]. Retrieved September 21, 2018 from http://arxiv.org/abs/1606.03490.

[53] Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible Models for Classification and Regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, 150–158. https://doi.org/10.1145/2339530.2339556

[54] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. 10.

[55] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. 2012. AffectAura: An intelligent system for emotional memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 849–858.

[56] B. Micenková, R. T. Ng, X. Dang, and I. Assent. 2013. Explaining outliers by subspace separability. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining*. 518–527. https://doi.org/10.1109/ICDM.2013.132

[57] Tim Miller. 2017. Explanation in artificial intelligence: Insights from the social sciences. arXiv:1706.07269 [cs]. Retrieved September 17, 2018 from http://arxiv.org/abs/1706.07269.

[58] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Dig. Sign. Process.* 73 (2018), 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

[59] Kathleen L. Mosier, Linda J. Skitka, Susan Heers, and Mark Burdick. 1998. Automation bias: Decision making and performance in high-tech cockpits. *Int. J. Aviat. Psychol.* 8, 1 (1998), 47–46. https://doi.org/10.1207/s15327108ijap0801_3

[60] Bonnie M. Muir and Neville Moray. 1996. Trust in automation. Part II. Experimental Studies of Trust and Human Intervention in a Process Control Simulation. *Ergonomics* 39, 3 (1996), 429–460. https://doi.org/10.1080/00140139608964474

[61] Saurabh Nagrecha, John Z. Dillon, and Nitesh V. Chawla. 2017. MOOC Dropout prediction: Lessons learned from making pipelines interpretable. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW Companion'17)*. 351–359. https://doi.org/10.1145/3041021.3054162

[62] Lloyd H. Nakatani and John A. Rohrlich. 1983. Soft machines: A philosophy of user-computer interface design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'83)*. 19–23. https://doi.org/10.1145/800045.801573

[63] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *J. Soc. Issues* 56, 1 (2000), 81–103. https://doi.org/10.1111/0022-4537.00153

[64] Cathy O'Neil. 2016. Weapons of math destruction: How big data increases inequality and threatens democracy. *Crown*.

[65] Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Books Limited.

[66] James W. Pennebaker. 2011. *The Secret Life of Pronouns: What Our Words Say about Us*. Bloomsbury Press.

[67] Richard E. Petty and John T. Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Advances in Experimental Social Psychology*, L. Berkowitz (Ed.), Vol. 19. Academic Press, San Diego, CA, 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2

[68] Pip. Retrieved September 21, 2018 from https://thepip.com/en-us/.

[69] Michael. F. Porter. 1980. An algorithm for suffix stripping. *Program* 14, 3 (1980), 130–137. https://doi.org/10.1108/eb046814

[70] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. 2015. MyBehavior: Automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 707–718.

[71] Lena Reed, Jiaqi Wu, Shereen Oraby, Pranav Anand, and Marilyn Walker. 2017. Learning lexico-functional patterns for first-person affect. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 141–147.

[72] Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers. In *Television, and New Media Like Real People and Places*. Cambridge University Press.

[73] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144. https://doi.org/10.1145/2939672.2939778

[74] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cogn. Emot.* 18, 8 (2004), 1121–1133. https://doi.org/10.1080/02699930441000030

[75] Emanuel A. Schegloff. 1992. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *Am. J. Sociol.* 97, 5 (1992), 1295–1345.

[76] David Canfield Smith. 1982. Designing the Star User Interface. *Byte Magazine*, April, 1982, 242–282.

[77] Aaron Springer and Henriette Cramer. 2018. "Play PRBLMS": Identifying and correcting less accessible content in voice interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. 296:1–296:13. https://doi.org/10.1145/3173574.3173870

[78] Aaron Springer, Victoria Hollis, and Steve Whittaker. 2017. Dice in the black box: User experiences with an inscrutable algorithm. Retrieved April 24, 2017 from https://aaai.org/ocs/index.php/SSS/SSS17/paper/view/15372.

[79] Aaron Springer, Victoria Hollis, and Steve Whittaker. 2018. Mood modeling: Accuracy depends on active logging and reflection. *Pers. Ubiquit. Comput.* 22 (2018), 723–737. https://doi.org/10.1007/s00779-018-1123-8

[80] Aaron Springer and Steve Whittaker. 2018. What are you hiding? algorithmic transparency and user perceptions. In *Proceedings of the 2018 AAAI Spring Symposium Series*.

[81] Litza Stark, Steve Whittaker, and Julia Hirschberg. 2000. ASR satisficing: The effects of ASR accuracy on speech retrieval. In *Processing of the International Conference on Speech and Language Processing (ICSLP-2000)*, 1069–1072.

[82] Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*. 53.

[83] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The illusion of control: Placebo effects of control settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. 1–16:13. https://doi.org/10.1145/3173574.3173590

[84] Jeffrey Warshaw, Tara Matthews, Steve Whittaker, Chris Kau, Mateo Bengualid, and Barton A. Smith. 2015. Can an algorithm know the "real you"?: Understanding people's reactions to hyper-personal analytics systems. 797–806. https://doi.org/10.1145/2702123.2702274

[85] Daniel S. Weld and Gagan Bansal. 2018. The challenge of crafting intelligible intelligence. arXiv:1803.04263 [cs]. Retrieved September 20, 2018 from http://arxiv.org/abs/1803.04263.

[86] Wickens Christopher, Hollands Justin, Banbury Simon, and Parasuraman Raja. 2015. Engineering Psychology and Human Performance, Prentice Hall, New Jersey.

[87] Jenna Wiens and Erica S. Shenoy. 2018. Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. *Clin. Infect. Dis.* 66, 1 (2018), 149–153. https://doi.org/10.1093/cid/cix731

[88] Woebot—Your Charming Robot Friend Who Is Here for You, 24/7. Retrieved September 21, 2018 from https://woebot.io.

[89] Rayoung Yang and Mark W. Newman. 2013. Learning from a learning thermostat: Lessons for intelligent systems for the home. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'13)*. 93. https://doi.org/10.1145/2493432.2493489

[90] Miriam Zisook, Sara Taylor, Akane Sano, and Rosalind Picard. 2016. SNAPSHOT Expose: Stage based and social theory based applications to reduce stress and improve wellbeing. Retrieved March 10, 2017 from https://pdfs.semanticscholar.org/b283/53899d0c5059a31c9bc69c364e62bc6c7ff5.pdf.

Received September 2019; revised December 2019; accepted February 2020