Perceive, Understand & Predict - Empirical Indication for Facets in Subjective Information Processing Awareness

TIM SCHRILLS, University of Lübeck, Germany
SUSANNE KARGL, University of Lübeck, Germany
MONA BICKEL, University of Lübeck, Germany
THOMAS FRANKE, University of Lübeck, Germany

When quantitatively measuring how humans experience the cooperation with intelligent systems, validated instruments grounded in established fields such as human-automation interaction research are needed. We derived Subjective Information Processing Awareness (SIPA) from core conceptions of Situation Awareness. SIPA describes to what extent interacting with an intelligent system enables users to experience 1) Transparency 2) Understandability and 3) Predictability of the system's information processing. The SIPA concept was operationalized as a 6-item scale. The goal of the present research was to examine the psychometric characteristics of the SIPA scale as well as the internal structure based on three samples from independent experiments with N = 162 participants. The SIPA scale showed possible applicability both as (a) a three-facet scale (three highly correlated facets) and (b) a one-dimensional scale as well as high reliability. Construct validity with related constructs such as explanation satisfaction or trust was expectedly high. Based on the results, the SIPA scale appears to be a promising tool for research on cooperative AI systems. We discuss what role SIPA and different effects of explanations on experienced traceability may have for future studies of, e.g., Human-AI teams.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **User studies**.

Additional Key Words and Phrases: human AI cooperation, XAI, traceability, user experience

ACM Reference Format:
Tim Schrills, Susanne Kargl, Mona Bickel, and Thomas Franke. 2022. Perceive, Understand & Predict - Empirical Indication for Facets in Subjective Information Processing Awareness. In . ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Quantifying the experience of users cooperating with intelligent systems is a challenging task and many different instruments emerge to support XAI research [41]. Empirically validated tools are crucial in order to 1) research explainable artificial intelligence (XAI) as well as 2) improve transparency for users of, e.g., machine learning models [10]. While tackling the disadvantages of opaque systems, many approaches aim to improve conditions for joint human-AI activity [23] by providing, e.g., visual explanations of the system's information processing [5]. In joint human-AI activity, at least one human and an autonomous system work together to complete a task [29] and explanations may help to enable, e.g., trust in the AI partner (cf. [40]). However, various studies [2, 11, 36] show that enriching the interaction with explanations can result in undesirable effects, e.g., complacency [31] or information overload, thus inhibiting the performance of human-AI cooperation. In particular, explanations that convince rather than provide helpful information for interpreting results are a potential hazard of AI systems [3].

Manuscript submitted to ACM

According to [23], successful cooperation between humans and machines requires 1) the goal to cooperate, 2) a common ground on which they operate, 3) mutual predictability, and 4) mutual directability. [7] extend this requirement by adding that cooperation needs collegiality that changes, e.g., in the objectives of one partner can be actively registered and included by the other partner. Since the actual effect of explanations on, e.g., the predictability of a system may differ from the perceived effect [1], the conditions for cooperation should be examined with both performance and experience measures. For this purpose, concepts from the field of human-automation research and engineering psychology can form the theoretical basis for appropriate measurement methods.

Situation Awareness (SA) [14] as a central concept is immensely associated with automation (cf. [30]) in human factors research. SA assumes three different levels of awareness to be relevant in a dynamic situation: through 1) perception of relevant stimuli and the subsequent 2) understanding of the underlying contexts, the 3) prediction of how the situation will develop is made possible. Based on these levels, we derive Subjective Information Processing Awareness (SIPA) [35], which can be defined as the experience of being enabled by a system to 1) perceive, 2) understand and, 3) predict its information processing. We assume that for a user to experience cooperation the different facets of SIPA must be addressed: a common ground (cf. [23]) requires, for example, that both partners have access to relevant input variables. Another core concept of SA projection is enabled by predictability, which is a central criterion for cooperativeness and SIPA. To allow for an assessment of SIPA experienced by users after interacting with an intelligent system, a highly economic scale was developed [35]. In contrast to other constructs SIPA is rooted in a psychological theoretical background of Human-AI-Cooperation. Further, it focuses on the subjective experience of a user who is interacting with an automated system rather than the understanding of explanations or provided information. Therefore, the construct is able to quantify the subjective traceability of Human automation Interaction.

The objective of the present work was to examine the dimensionality and validity of the SIPA scale in the context of three different samples (S1 - S3). To this end, SIPA data from independent experiments conducted in two different contexts (Automated Insulin Therapy and EcoCharging) was analysed.

2 COOPERATION OF HUMAN-AI TEAMS AND SUBJECTIVE INFORMATION AWARENESS

A theoretical approach to describe the development of awareness for information processing when cooperating with intelligent systems (or automation in general) may be derived from SA. Based on SA, [4] establish the concept of situation models, which - related to information processing - could be interpreted similarly to mental models. The latter is repeatedly discussed in the field of human-AI cooperation [34]. The development of situation awareness, which is shown by [17] in the presentation of SA, can be transferred to information processing and can be supported by system design (cf. also [16]). In the first step, (SIPA-1) transparency of the system enables users to (SA-1) *perceive* relevant information of information processing. In the second step, the (SIPA-2) understandability of a system leads to a correct (SA-2) *understanding* of the system's information processing. Mental models, which are important for understanding in Human-AI Interaction [25], describe the possibility to perform mental manipulations, which, accordingly, could be achieved via different interaction possibilities with the system (see also [33], [26] on counterfactual explanations). Finally, systems that exhibit high (SIPA-3) predictability allow for (SA-3) *projection* of future processing to be captured as a function of the given information. Predictability plays an important role for a user to regulate his or her own actions during cooperation with an automated system. If a system is highly predictable it allows the user to more accurately estimate the effect of their own behaviour on the system. This enables them to manage their contribution in the interaction process and therefore improve human automation interaction.

While objective measurement methods for capturing SA (see [15]) may be applied to the use of intelligent automation in collaborative scenarios, fewer such methods exist for experienced SA (for example, SART [37]). However, the matching of objective and subjective metrics of interaction is important in order to detect, e.g., placebic explanations (as described in [11]) or even dark patterns [8]. To develop a scale able to measure SIPA as a subjective awareness construct, 12 items were first generated based on existing literature on SA. These were tested in a study with two different systems [35] and each item was examined with regard to item difficulty and dispersion. In combination with qualitative comments, the scale was condensed to 6 items. The items of the scale were developed in parallel for German and English language.

The English SIPA scale is shown in Table 1. The SIPA scale consists of six items and uses a 6-point Likert scale. When entering participants' responses in a data file for the analysis, the responses are coded as follows: completely disagree = 1, largely disagree = 2, slightly disagree = 3, slightly agree = 4, largely agree = 5, completely agree = 6. Finally, a mean score is computed over all six items to obtain a person's SIPA score (i.e., 1.0-6.0). There is the option to calculate the mean value individually for transparency (T-1 & T-2), understandability (U-1 & U-2) and predictability (P-1 & P-2). Whether the mean value should be calculated for all items or for the facets depends on the research question that is to be investigated.

Table 1. All Items of the Subjective Information processing (SIPA) Scale and the corresponding instruction

The following questionnaire deals with your **experience in the interaction with the system.** **Information** refers to all data that the system can work with. **Result** refers to the output of the system, which is presented at the end of the system's information processing

Please indicate the degree to which you agree/disagree with the following statements		completely disagree	largely disagree	slightly disagree	slightly agree	largely agree	completely agree
T-1	It was transparent to me which information was collected by the system.						
T-2	The information that the system could acquire was observable for me.						
U-1	It was understandable to me how the collected information led to the result.						
U-2	The system's information processing was comprehensible to me.						
P-1	With the information accessible for me, the results was foreseeable for me.						
P-2	The system's information processing was predictable for me.						

3 EMPIRICAL APPLICATIONS OF SIPA IN HUMAN-AI COOPERATION

We examined the dimensionality, reliability and validity of the SIPA scale in the context of three different samples (S1 - S3): in S1 the scale was used in the context of EcoCharging; S2 examined SIPA of participants with diabetes mellitus type 1 (DMT1) attempting to trace an algorithm for automated insulin delivery; and S3 evaluated SIPA after observation and interaction with a simulation of an automated insulin delivery (AID) system. Fig 1 depicts an example of one of the stimuli shown to participants in S2. The representations of the other experiments were analogous.

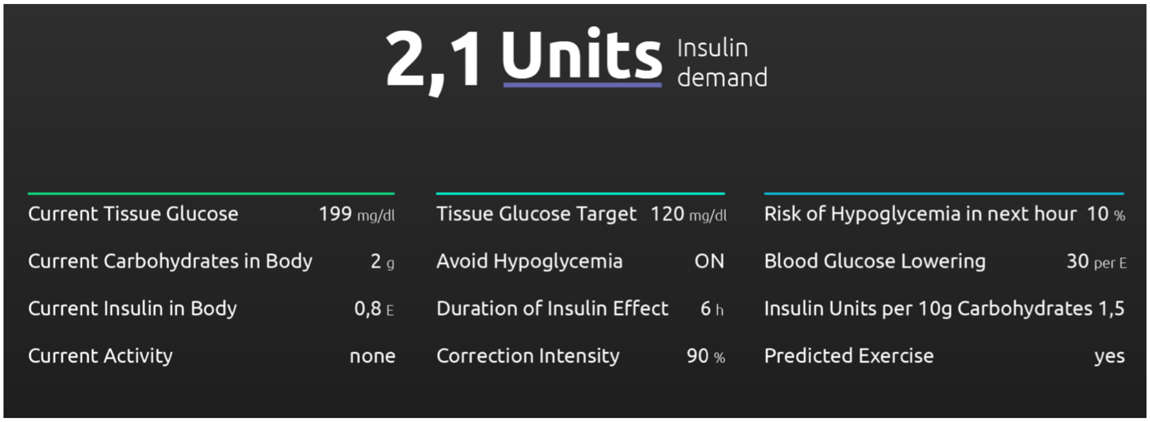

Fig. 1. Stimuli from the diabetes-related experiment (S2) as they were shown to participants.

In the first sample (S1) participants were shown estimates of eco-friendliness in form of tokens for different bookings of electrical cars with different levels of information disclosure, providing details about data that was used by an EcoCharging algorithm to calculate these prices (often labelled as Smart charging or intelligent charging [19]). The instructed task was to try to trace how the algorithm calculated the number of tokens. In five similar observation blocks with 10 stimuli each, users had to observe and draw conclusions about the underlying algorithm. After each block SIPA and trust (using the FOST Scale [38]) were measured.

The experiment for S2 was constructed similarly. Here, participants had to observe the calculations for insulin requirement of an automated insulin delivery algorithm. Hence, only people with DMT1 participated in the experiment. Users saw four similar blocks of 15 stimuli each, and were asked to observe and draw conclusions about the underlying algorithm. After each block SIPA and trust (using the FOST Scale [38]) were measured. Finally, the automated insulin delivery simulation used to create scenarios for S2 was transformed into an interactive simulator for S3.

Participants were also asked to complete the Explanation Satisfaction Scale (ESS) in S1 and S2 [20] to explore the scale's construct validity. Additionally, to subjective measures regarding experienced interaction, participants' Affinity for Technology Interaction (ATI) was assessed [18].

For the purpose of the following analyses only the values from the last time the participants filled out the SIPA questionnaire for each experiment are used. This refers to the fifth observation in S1, the fourth in S2 and the first in S3. S3 was analysed separately due to its different structure and experimental design.

4 RESULTS

The sample sizes of S1-S3 were $n_{S1} = 60$, $n_{S2} = 70$ and $n_{S3} = 32$ participants. Age and gender were similarly distributed among the different samples as shown in Table 2. In all samples multiple SIPA measures were obtained. Since not all variables showed normal distribution and linearity, Spearman's Rho was calculated for correlations and is depicted as r_S; p-values for all test families (i.e., for each type of analysis) were corrected using the Bonferroni-Holm method [21].

4.1 Dimensionality

According to the dimensions of SA we assumed three correlated factors: transparency (T-1 and T-2), understandability (U1- and U-2) and predictability (P-1 and P-2). To verify this assumption we carried out a confirmatory factor analyses

Table 2. **Descriptive Data of S1-S3.**

Sample	***n***	***Mean* Age (*SD*)**	**Gender (f:m:d)**
S1	60	24.1 (9.26)	43:15:2
S2	70	28.9 (10.5)	49:20:1
S3	32	32.3.1 (11.1)	not reported

Table 3. **Component loadings.**

	Component			
	1	2	3	Uniqueness
SIPA T-1	.91			.06
SIPA T-2	.99			.05
SIPA U-1			.82	.03
SIPA U-2			.96	.03
SIPA P-1		.99		.03
SIPA P-2		.83		.04

Note. 'promax' rotation was used.
Loadings <.3 are hidden

with the program jamovi [22]. We compared a one-factor model (χ^2 = 193, p < .001, *RSME* = .396, (90% CI: .348, .446), *CFI* = .798, *TLI* = .663, *SRMR* = .076) with a three-factor model (χ^2 = 8.24, p = .221, *RSME* = .054, (90% CI: .000, .143), *CFI* = .998, *TLI* = .994, *SRMR* = .010). The resulting model with factors transparency, understandability and predictability had an overall better model fit than the model that assumed a one factor solution. Additionally, a principal component analysis with three forced factor was conducted, to estimate each items' contribution to the different facets. Table 3 shows the factor loadings resulting from a principle component analysis with three components as well as each item's uniqueness, which represents the variance that is 'unique' to the variable and not shared with other variables. All in all, the assumption of three latent facets in SIPA similar to the levels from SA can be supported.

The facets transparency, understandability and predictability weren't independent of each other. Transparency correlated moderately and significantly with understandability in all three samples (S1: r_S = .58, p < .001, S2: r_S = .69; p < .001, S3: r_S = .42, p = .017). However, the correlation of predictability and transparency only reached significance in S1 (r_S = .50; p = .004) and S2 (r_S = .67; p < .001) but not in S3 (r_S = .35; p = .053). Predictability and understandability correlated strongly with each other and reached significance in all three samples (S1: r_S = .86, p < .001, S2: r_S = .84; p < .001, S3: r_S = .75, p < .001).

4.2 Reliability

McDonald's omega (ω) is used to assess reliability in addition to Cronbach's α, since the latter is not well-suited for short scales [9]. Internal consistency of the entire scale was high in all three samples (S1: McDonald's ω = .93, Cronbach's α =.92; S2: McDonald's ω = .94, Cronbach's α = .90; S3: McDonald's ω = .86 Cronbach's α = .84). Since the facets only consisted of two items each, the Spearman-Brown coefficient was used to assess their reliability [13]. The items of the transparency facet showed moderate to high consistency in all three samples (S1: R = .64, S2: R = .74, S3: R = .54). The U-1 and U-2 that form the understandability facet showed high consistency in all three samples (S1: R = .91, S2: R = .83, S3: R = .76). The items of the predictability facet all showed high consistency in all three samples (S1: R = .92, S2: R = .81, S3: R = .93).

4.3 Construct Validity

To evaluate the construct validity of the scale we examined correlations with other related questionnaires in the three samples. In all three samples the total SIPA score correlated moderately to strongly with trust (S1: r_S = .84, p = .002, S2: r = .86; p = .001, S3: r_S = .55, p = .016). All sub-scales of SIPA also correlated significantly with FOST in S1 (transparency:

r_S = .47; p = .028/ understandability: r_S = .85; p = .001/ predictability: r_S = .75; p = .000) and in S2 (transparency: r_S = .65; p = .018/ understandability: r_S = .84; p = .002/ predictability: r_S = .79; p = .003). However, in S3 only understandability (r_S = .58; p = .021) and predictability (r_S = .55; p = .023) but not transparency (r_S = .14; p = .433) reached significance.

In contrast, all correlations between affinity for technology and SIPA (total or as facets) were not significant (all p > .050). The explanation satisfaction scale was only completed by participants in S1 and S2. Therefore, we calculated the correlations of the ESS and the SIPA only in these samples. The total SIPA score as well as understandability and predictability correlated moderately with the ESS in S1 (Total SIPA: r_S = .55, p = .018; understandability: r_S = .55; p = .023; predictability: r_S = .79; p = .009). However, the correlation with transparency did not reach significance (r_S = .31; p = .121) The total scale as well as all facets correlated moderately to strongly with the ESS in S2 (Total Score: r_S = .70, p = .008; transparency: r_S = .57; p = .020/ understandability: r_S = .67; p = .013; predictability: r_S = .66; p = .013).

Table 4. Descriptive Data

	M			*SD*			Kolmogorov-Smirnov Test		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
SIPA T-1	4.08	4.64	4.00	1.21	1.19	1.39	p = .005*	p <.001*	p = .211
SIPA T-2	3.88	4.67	4.31	1.25	1.24	1.20	p = .006*	p <.001*	p = .061
SIPA U-1	3.07	3.90	3.09	1.23	1.40	1.40	p = .050*	p = .001*	p = .036*
SIPA U-2	2.88	3.86	2.94	1.25	1.44	1.39	p = .117	p = .031*	p = .006*
SIPA P-1	2.95	3.66	2.81	1.21	1.38	1.33	p = .105	p = .036*	p = .009*
SIPA P-2	2.72	3.56	2.78	1.21	1.40	1.21	p = .049*	p = .040*	p = .004*

Additionally, we carried out item analyses in all three samples. Table 4 shows mean, standard deviation, as well the p-value of the Kolmogorow-Smirnow-Test, which was used to check the distribution of the items. Mostly the test was significant indicating that the items are not normally distributed.

To assess whether SIPA is able to discriminate across different user experiences of human-AI interaction, item difficulty and item discrimination values (according to [27]) were calculated (see Table 5). Values of item difficulty ranged on average from 46.7% (SIPA P-2) to 71.9% (SIPA T-2). These are satisfactory results given that moderate item difficulties (i.e., 50%) can differentiate best between participants with high and low SIPA after an interaction ([27]). However, items of the transparency facet (SIPA T-1 & T-2) need to be monitored, since they reached high values in S2 (i.e. many participants rated the item high). Values of item discrimination (i.e., part-whole corrected item-total correlations) ranged on average from = .60 (SIPA T-2) to .85 (SIPA U-2). These results indicate good item discrimination [27]. However, item discrimination is quite high for SIPA U-1, U-2, P-1 and P-2, which means that there may be room to reduce the number of items in the future. In sum, the SIPA items are able to differentiate between high- and low-SIPA participants, allowing them to be used in human-AI interaction studies with various designs or explanations to be tested.

5 DISCUSSION

The aim of the present work was to assess the quality of the SIPA scale, which measures users' experience of interacting with intelligent automation, e.g., in human-AI cooperation. Overall, the scale showed very good construct validity as correlations with existing measures of human-technology interaction, such as those used to assess satisfaction with explanations and trust, were expectedly high. **On top of a high internal reliability, the empirical results allow**

Table 5. Item Analysis Results

	Item difficulty				Item discrimination			
	S1	S2	S3	*M*	S1	S2	S3	*M*
SIPA T-1	68.3	71.9	66.7	69.0	.70	.85	.52	.69
SIPA T-2	65.8	74.3	71.9	70.7	.73	.76	.32	.60
SIPA U-1	55.3	58.8	51.6	55.2	.90	.90	.64	.81
SIPA U-2	52.2	58.3	49.0	53.2	.84	.87	.83	.85
SIPA P-1	50.3	55.0	46.9	50.7	.82	.84	.77	.81
SIPA P-2	46.7	53.3	46.4	48.8	.82	.86	.76	.81

conclusions about three highly correlated SIPA facets, which can be based on the scale's theoretical foundation. The application of the **SIPA scale supports research of explanations in intelligent automation** and is suitable for use, for example, to address the following research issues empirically.

5.1 Use SIPA to evaluate different explanation approaches

[24] compared different explanatory methods (e.g., Lime, SHAPE, and CUI) and found that they have different effects on experienced transparency, understandability, or satisfaction. Overall, several studies indicate a close connection between the function of an explanation and the situation in which it is given, e.g., [5]. However, the results of the present study indicate that diverging effects of explanations on experience could affect different facets of SIPA. SIPA can be used to evaluate explanations' effect of experienced traceability in more detail in different contexts (e.g, EcoCharing or AI-based diagnosis).

5.2 Use SIPA to evaluate the relationship between experienced traceability and performance

[3] report that explanations in a machine learning context increase the chance that an incorrect output result of a model will be incorrectly accepted by the user. One reason may be that explanations do not add objective value to the traceability, but rather function as a placebo [12], possibly because they are improperly processed due to insufficient resources, like workload [32]. Placebic explanations may, in the worst case, even be consciously intended to mask the low performance of a system, i.e., of dark patterns [8]. Low levels of trust are not necessarily equal to poor traceability of the system. For example, in case of a system users consider inadequate for a specific task, an increasing SIPA value may enable a decrease in trust. In addition to objective measures, the use of the SIPA scale is suitable for identifying explanations that give users a valid feeling of better traceability. For example, if only the trust value is observed, one could conclude that a particular explanation weakens or even changes the relationship between performance (e.g., a prediction) and users' experience. SIPA thus represents an important complement to, e.g., performance measures in human-AI cooperation. Further studies need to examine the relationship between SIPA and trust or the perception of cooperation (e.g., whether SIPA poses a prerequisite for trust as perceived predictability may be a central criterion for trust). SIPA can be used to identify situations where explanations are confusing or not yet optimal and therefore can negatively affect a user's performance. The scale can therefore be used to explore how a user's interaction with an automated system can be optimized and which features support or hinder the resulting performance.

5.3 Use SIPA to evaluate the effect of interactions on experienced traceability

Several studies report a positive effect of interactive explanations on comprehensibility and user experience, e.g., simulations, which offer users the possibility to explore the functions of AI systems [6, 28]. That means, SIPA values should show higher correlations with performance after an interactive exploration period than after an observation period. In future studies, different explanatory approaches should be systematically examined for their effects on objective and subjective interaction measures, i.e., collaborative performance or SIPA. For this purpose, a setup where users can systematically vary which questions they ask to a system (e.g., "What if?" versus "Why not x?" as described in [39]), should be examined in more depth in the future. The SIPA scale, through its different facets, can help to shed light on the effects of explanations while being based on theoretical foundations in Human Factors research.

6 CONCLUSION

The present work examined the SIPA scale in terms of dimensionality, reliability and validity. To this end, the results of three independent experiments from the fields of Automated Insulin Delivery and EcoCharging were analysed. While the scale overall demonstrated very good reliability and convergent validity, it also showed three highly correlated facets that correspond to the theoretical levels from SA. Additionally, our findings indicate that it is advantageous to use differentiated methodologies (e.g. SIPA as well as trust measures) for the examination of subjective experience in human-AI cooperation in order to investigate the context- and task-dependent effects of, e.g., static and interactive explanations in more depth.

ACKNOWLEDGMENTS

This research has been funded by the Federal Ministry of Education and Research of Germany in the framework of the project CoCoAI (Cooperative and Communicating AI, project number 01GP1908).

REFERENCES

[1] Yasmeen Alufaisan, Laura R. Marusich, Jonathan Z. Bakdash, Yan Zhou, and Murat Kantarcioglu. 2020. Does Explainable Artificial Intelligence Improve Human Decision-Making? *CoRR* abs/2006.11194 (2020). arXiv:2006.11194 https://arxiv.org/abs/2006.11194

[2] Gagan Bansal. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. (2019), 10.

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *arXiv:2006.14779 [cs]* (Jan. 2021). http://arxiv.org/abs/2006.14779 arXiv: 2006.14779.

[4] M. Baumann and J. F. Krems. 2007. Situation Awareness and Driving: A Cognitive Model. In *Modelling Driver Behaviour in Automotive Environments*, P. Carlo Cacciabue (Ed.). Springer London, London, 253–265. https://doi.org/10.1007/978-1-84628-618-6_14

[5] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–24. https://doi.org/10.1145/3359206

[6] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300789

[7] Erin K. Chiou and John D. Lee. 2021. Trusting Automation: Designing for Responsivity and Resilience. *Human Factors: The Journal of the Human Factors and Ergonomics Society* (April 2021), 001872082110099. https://doi.org/10.1177/00187208211009995

[8] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. 2019. Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems. In *Los Angeles*. 6.

[9] Thomas J Dunn, Thom Baguley, and Vivienne Brunsden. 2014. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British journal of psychology* 105, 3 (2014), 399–412.

[10] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O. Riedl. 2021. *Operationalizing Human-Centered Perspectives in Explainable AI*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411763.3441342

[11] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–6. https://doi.org/10.1145/3290607.3312787

[12] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–6. https://doi.org/10.1145/3290607.3312787

[13] Rob Eisinga, Manfred te Grotenhuis, and Ben Pelzer. 2013. The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health* 58, 4 (Aug. 2013), 637–642. https://doi.org/10.1007/s00038-012-0416-3

[14] M.R. Endsley. 1988. Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference*. IEEE, Dayton, OH, USA, 789–795. https://doi.org/10.1109/NAECON.1988.195097

[15] Mica R Endsley. 2011. Direct Measurement of Situation Awareness: Validity and Use of SAGAT. In *Situational Awareness*. Vol. 1. Taylor and Francis Group, 28. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315087924-9/direct-measurement-situation-awareness-validity-use-sagat-mica-endsley

[16] Mica R Endsley. 2012. *Designing for Situation Awareness: An Approach to User-Centered Design, Second Edition* (2 ed.). Taylor & Francis Group. https://www.routledge.com/Designing-for-Situation-Awareness-An-Approach-to-User-Centered-Design/Endsley/p/book/9781420063554

[17] Mica R. Endsley. 2015. Situation Awareness Misconceptions and Misunderstandings. *Journal of Cognitive Engineering and Decision Making* 9, 1 (March 2015), 4–32. https://doi.org/10.1177/1555343415572631

[18] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction* 35, 6 (April 2019), 456–467. https://doi.org/10.1080/10447318.2018.1456150 Number: 6.

[19] Javier García-Villalobos, Inmaculada Zamora, José Ignacio San Martín, Francisco Javier Asensio, and Víctor Aperribay. 2014. Plug-in electric vehicles in electric distribution networks: A review of smart charging approaches. *Renewable and Sustainable Energy Reviews* 38 (2014), 717–731.

[20] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for Explainable AI: Challenges and Prospects. *arXiv:1812.04608 [cs]* (Feb. 2019). http://arxiv.org/abs/1812.04608 arXiv: 1812.04608.

[21] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70. http://www.jstor.org/stable/4615733

[22] The jamovi project. 2021. *jamovi*. Retrieved from https://www.jamovi.org. Version 1.6 [Computer Software].

[23] G. Klein, D.D. Woods, J.M. Bradshaw, R.R. Hoffman, and P.J. Feltovich. 2004. Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity. *IEEE Intelligent Systems* 19, 06 (Nov. 2004), 91–95. https://doi.org/10.1109/MIS.2004.74

[24] Samanta Knapič, Avleen Malhi, Rohit Saluja, and Kary Främling. 2021. Explainable Artificial Intelligence for Human Decision-Support System in Medical Domain. *arXiv:2105.02357 [cs]* (May 2021). http://arxiv.org/abs/2105.02357 arXiv: 2105.02357.

[25] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, San Jose, CA, USA, 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

[26] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[27] Helfried Moosbrugger and Augustin Kelava. 2012. Testtheorie und Fragebogenkonstruktion. (2012).

[28] An T Nguyen, Matthew Lease, and Byron C Wallace. 2019. Mash: software tools for developing interactive and transparent machine learning systems. *Los Angeles* (2019), 7.

[29] Thomas O'Neill, Nathan McNeese, Amy Barron, and Beau Schelble. 2020. Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors: The Journal of the Human Factors and Ergonomics Society* (Oct. 2020), 001872082096086. https://doi.org/10.1177/0018720820960865

[30] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30, 3 (2000), 286–297. https://doi.org/10.1109/3468.844354

[31] Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens. 2008. Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making* 2, 2 (June 2008), 140–160. https://doi.org/10.1518/155534308X284417

[32] Arun Rai. 2020. Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science* 48, 1 (Jan. 2020), 137–141. https://doi.org/10.1007/s11747-019-00710-5

[33] William B. Rouse and Nancy M. Morris. 1986. On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin* 100, 3 (1986), 349–363. https://doi.org/10.1037/0033-2909.100.3.349

[34] Heleen Rutjes, Martijn Willemsen, and Wijnand IJsselsteijn. 2019. Considerations on explainable AI and users' mental models: CHI 2019 Workshop. *Where is the Human? Bridging the Gap Between AI and HCI* (May 2019). Place: New York, NY, USA Publisher: Association for Computing Machinery, Inc.

[35] Tim Schrills, Mourad Zoubir, Mona Bickel, Susanne Kargl, and Thomas Franke. 2021. Are Users in the Loop? Development of the Subjective Information Processing Awareness Scale to Assess XAI. (May 2021), 9.

[36] Aaron Springer and Steve Whittaker. 2018. "I had a solid theory before but it's falling apart": Polarizing Effects of Algorithmic Transparency. *CoRR* abs/1811.02163 (2018). arXiv:1811.02163 http://arxiv.org/abs/1811.02163

[37] R. M. Taylor. 2011. Situational Awareness Rating Technique (Sart): The Development of a Tool for Aircrew Systems Design. In *Situational Awareness*. Routledge. Num Pages: 18.

[38] Daniel Trommler, Christiane Attig, and Thomas Franke. 2018. Trust in activity tracker measurement and its link to user acceptance. (2018). https://doi.org/10.18420/MUC2018-MCI-0361 Publisher: Gesellschaft für Informatik e.V..

[39] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[40] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 295–305. https://doi.org/10.1145/3351095.3372852

[41] Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics* 10, 5 (March 2021), 593. https://doi.org/10.3390/electronics10050593 Number: 5.