How do Users Experience Traceability of AI Systems? Examining Subjective Information Processing Awareness in Automated Insulin Delivery (AID) Systems

TIM SCHRILLS, Universität zu Lübeck, Germany
THOMAS FRANKE, Universität zu Lübeck, Germany

When interacting with artificial intelligence (AI) in the medical domain, users frequently face automated information processing, which can remain opaque to them. For example, users with diabetes may interact daily with automated insulin delivery (AID). However, effective AID therapy requires traceability of automated decisions for diverse users. Grounded in research on human-automation interaction, we study Subjective Information Processing Awareness (SIPA) as key construct to research users' experience of explainable AI. The objective of the present research was to examine how users experience differing levels of traceability of an AI algorithm. We developed a basic AID simulation to create realistic scenarios for an experiment with $N = 80$, where we examined the effect of three levels of information disclosure on SIPA and performance. Attributes serving as basis for insulin needs calculation were shown to users, who predicted the AID system's calculation after over 60 observations. Results showed a difference in SIPA after repeated observations, associated with a general decline of SIPA ratings over time. Supporting scale validity, SIPA was strongly correlated with trust and satisfaction with explanations. The present research indicates that the effect of different levels of information disclosure may need several repetitions before it manifests. Additionally, high levels of information disclosure may lead to a miscalibration between SIPA and performance in predicting the system's results. The results indicate that for a responsible design of XAI, system designers could utilize prediction tasks in order to calibrate experienced traceability.

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

Additional Key Words and Phrases: explainability, trust, human-centered AI, human AI cooperation

ACM Reference Format:
Tim Schrills and Thomas Franke. 2022. How do Users Experience Traceability of AI Systems? Examining Subjective Information Processing Awareness in Automated Insulin Delivery (AID) Systems. In . ACM, New York, NY, USA, 32 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The availability of intelligent technology for type 1 diabetes mellitus (DMT1) therapy [33] increases, reflecting the general development of personalized medicine based on artificial intelligence (AI). In DMT1, self-adapting learning algorithms are used for personalized calculation of insulin needs, e.g., at different times of the day, at different stages of the female period, or depending on physical activity. The goal of these systems, also known as automated insulin delivery (AID) systems, is to improve therapy while reducing the workload for people with DMT1. The incidence of DMT1 has increased in recent years and was 15 per 100,000 cases in 2020 [82]. In order to improve therapy conditions and effectiveness, AID systems can provide fully or partially automated diabetes therapy, for example, through integrating advanced wearable glucose sensors and intelligent insulin pumps [115]. All in all, the core of AID technology is the

Manuscript submitted to ACM

automated processing of information, especially to regulate current blood glucose levels in relation to therapy goals while dealing with high temporal dynamics, latency and complexity of human physiology.

First empirical studies suggest that people with DMT1 can benefit significantly from AID systems [3, 18, 63]. Both long-term metrics (e.g., the "time in range" (TIR) referring to desired glucose level) and the frequency of acute life-critical blood glucose levels can be reduced [6]. However, the positive effect of AID systems seems to depend on, for example, the previous quality of therapy [15, 79]. That is, individuals who had problematic long-term metrics before starting AID therapy are more likely to discontinue AID based therapy. Paradoxically, they would profit the most from AID systems. Thus, more inclusive methods that enable a wide diversity of users to continue AID therapy are needed. Parallel to findings on the beneficial therapuetic effects of AID therapy, several recent studies [4, 40, 79] explicate the need for human-centered development of AID systems, referring to problems well known in human-automation interaction: positive effects of AID can, e.g., be hindered by a high number of alarms [14] and the associated alarm fatigue [105]. While reducing the burden of treatment [112] is one of the main goals of AID systems, the continuous efforts while using AID systems as well as initial familiarization with this form of therapy are considered important discontinuation criteria for therapies with AID systems [79]. Human-centered improvement of the interaction between intelligent, highly adaptive AID systems and people with DMT1 is therefore a key scientific challenge to improve treatment options for individuals with different levels of experience and competence in using technology. At the same time AID systems also provide an excellent context to examine the dynamics of human-XAI interaction in a situation where high risks and high benefits for users are juxtaposed.

Problematic expectations and experiences with AID systems play a decisive role in the current acceptance of these systems [71]. For instance, if users have an incorrect understanding (e.g., in the sense of an inaccurate mental model, c.f. [58]), this can lead to incorrect predictions of the results and capability of the system [9]. Such false mental models could result from people being uncertain how system adaptability affects information processing in AID systems, e.g., whether they are able to change therapy goals or not [66]. In addition, AID systems often work differently than users did when they manually regulated their glucose levels: for example, information is processed by AID systems every 5 minutes [12], while in other forms of therapy (e.g., before using an AID system) the blood glucose level is sometimes only checked e.g four times a day with fingerstick glucose measurements [119]. Therefore, AID systems as a case for examining the real-time cooperation of humans with intelligent algorithms potentially lead to an advanced understanding of cooperative disease management between humans and AI. The performance of many AID systems regularly relies on information from the user [20, 115], so correct communication between both partners may lead to increased performance. On the other side, an incorrect understanding of the AID system could also have a critical impact on the success of the therapy [21]. While regulatory technical briefing is mandatory, the extent to which the functions and capabilities of such a system are understood is not tested prior to its use. If users have an incorrect mental model, the ability to correctly predict the information processing of the system may decrease. However, the self-assessment of how well one understands the information processing of a system may differ from the actual correctness. Explanations could help individuals to recognize errors in their mental model, leading to a better fit between experienced traceability and performance. However, they could also erroneously increase the confidence in an incorrect mental model and thus worsen the calibration [34], which results in wrong expectations about system behavior and potentially confuses users, ultimately leading to a reduction of trust [109]. Explanations can have an ambiguous effect on the calibration between experienced traceability of a system and the user's ability to correctly predict information processing. To address inaccurate calibration, metrics for both experience and performance need to be measured at the same time. All in all, AID systems represent a prototypical example of interactive systems where human-machine cooperation is centrally influenced by user experience and where incorrect mental models or disparity between experienced traceability and performance may lead to unexpected issues in therapy quality.

The goal-oriented communication of information, as well as the correct predictability of, e.g., an insulin calculation are two central characteristics of human-machine-cooperation [61]. In the field of explainable AI (XAI), various approaches exist that are intended to help users cooperate with AI systems by addressing the challenge of opacity (such as [25, 83, 95]. As demonstrated in examples outside of AID therapy the calculation of results can be presented transparently by revealing weights of relevant factors [99]. Furthermore, the elements that particularly favored certain results can be highlighted [69], or alternatives close to the given result can be presented [23]. In addition to improving predictability, explanations in AID systems could also help improve users' opportunities to exert directability (see [57] and [28]) . In DMT1, a loss of 'sense of control' is a typical problem users experience [104]. Thus, when using intelligent AID systems, increasing directability could play an important role and influence acceptance. Ultimately, "common ground" is an important prerequisite for cooperation [61]. In the case of AID systems, a common ground could consist of 1) current information on blood glucose levels, physical activity or food intake, 2) reference values for therapy, i.e., goals, or 3) personalized parameters like insulin sensitivity. Therefore, it is important to disclose relevant elements or information that users can process themselves and use to manually adjust the therapy [94], see also [110]. However, in order to reduce the workload, many AID systems process information automatically and do not actively share it with the users. This barriers have already led to user-initiated projects enabling access to their data (cf. [96]). Yet, in relation to the clinical relevance and the opportunities for human factors research, empirical studies on how and when to present detailed information on the AID's information processing is still in a early stage of development. Comprehensive and empirical work with a high ecological validity to derive guidelines on how AID systems can be improved to enable cooperation is needed and constitutes an important next step in human-centered diabetes technology.

The objective of the present research was to examine the effects of explanations that vary in the amount of disclosed information as well as repeated interaction on users' subjective perception of trust and traceability in AID systems. To this end, we trained a basic, yet prototypical AID algorithm based on artificial yet plausible data and designed a minimalistic AID simulation to create stimuli for an online experiment, where people with DMT1 repeatedly interacted with AID calculations and also predicted AID results. The information available to the algorithm was disclosed to participants to a different extent, in order to create three different experimental conditions. It was investigated whether a greater amount of information leads to higher experienced traceability and trust, while task completion time and perceived workload increase. Furthermore, it was analyzed to what extent repeated viewing of explanatory information can lead to an increase in experienced traceability. Similarly, the relationship between experienced traceability and the ability to make correct productions was assessed to allow evaluation of the calibration of the mental model with the system's information processing.

2 RELATED WORK

2.1 Automation in Diabetes Mellitus Typ 1

The continuous therapy of DMT1 sometimes can represent a great burden in everyday life for those affected [111]. Many therefore expect the digitalization of diabetes therapy to improve the quality of treatment while at the same time reducing the burden of treatment for patients [67]. This goal is also being pursued by the development of an "artificial pancreas", which allows complete automation of diabetes management [115]. For now, full automation is only possible to a limited extent due to various factors or may be associated with reduced precision of the therapy (c.f. [20]).

AID systems in the form of so-called hybrid closed-loop systems acknowledge those limits, while still offering relief for patients. These systems are not fully automated, since a system-dependent level of information or decisions by the user is required. [88] provide a suitable framework that distinguishes four stages of information processing (1. information acquisition, 2. information analysis, 3. decision making & 4. action implementation) and therefore allows a characterization of AID systems' level of automation. For example, there are already differences between existing systems in **information acquisition** (1): the system described by [12] only requires information on physical activity and food intake, while [45] already no longer requires information on physical activity. In **information analysis** (2), AID systems show a high degree of automation, as this is supposed to be a crucial element of relief for the users. Here, learning systems such as [12] can be distinguished from static systems such as [19]; the latter requires users to manually adjust parameters and thereby increase the quality of information analysis, whereas this is not necessary for self-learning systems. Thus, self-learning AID systems promise continuous improvement in therapy with greater automation, yet may be more complex to understand and to predict for users. The (3) **decision making** of, e.g., administration of insulin can be illustrated very well by the levels of automation presented by [88] and at the same time represents an important feature for interaction design in AID systems. For example, after input, a single suggestion for the administration of insulin can be made (level 4 cf. [91]) or an automatic administration of insulin occurs where the user can intervene but is not informed in any case (level 8). **Action Implementation** (4) is performed automatically by many systems in the event of identified insulin needs. However, systems currently available do not offer the injection of, e.g., glucose in case of hypoglycemia, so action implementation for low glucose level is not automated. All in all, AID systems in their various forms represent not only a broad field of automation in medical systems, but also systems that are highly dependent on cooperation between humans and technology.

However, various studies also show the challenges of automation: for example, people fear an error-proneness of digital systems in the field of DMT1 with simultaneous fears to be faced with high complexity [79]. But also, for example, too high expectations of performance or degree of system autonomy, especially of AID systems without a high degree of automation, pose substantial challenges [60, 92]. Furthermore, it remains to be seen to what extent a more technologized therapy could further exacerbate the already existing inequality between individuals from different socioeconomic strata or educational levels. In addition to accessibility (c.f. [68]), the design of systems may also improve unequal opportunities for empowered and autonomous diabetes therapy [73, 86]. These challenges can be addressed with the human-centered development of interactive and cooperative yet traceable AID systems, which could make a decisive contribution to the empowerment of people with DMT1, regardless of their diverse backgrounds, e.g., in terms of affinity to technological interaction or educational level.

2.2 Explanation and Cooperation in AID systems

Explanations and higher levels of transparency may improve cooperation between humans and intelligent systems [117]. They may support the temporally adequate exchange of information between humans and the system, which is of central importance for both partners to fulfill their respective functions [47]. In AID systems, for example, the human must signal the intake of carbohydrates timely, while the system must communicate a deviation in blood glucose levels to the user, for example, so that the human can take action. Mutual anticipation of information demands can be a central criterion of cooperation in the sense of collegiality (cf. [28]). Especially with higher degrees of automation, the human's task can also be to monitor or check results. For this task, the information used by the machine can be a central function for cooperation, as this allows the inputs for the machine calculation to be traced. The extent to which the information processing of a system is accessible for the users and thus also provides the basis for cooperative actions can be described as traceability (unlike the definition of [65], where traceability refers to the creation process of the system and not of an individual calculation). An empirical investigation of the disclosure of information in the context of a decision-making process can therefore make an important contribution to the design of human-centered AID systems. To the best of our knowledge, no results on how different quantities of information contributed to the calculation of insulin needs affect user experience have been published.

However, communication - if it does not take place at the right time - can have negative effects on cooperation or the performance of other functions by a partner [32]. Accordingly, previous research does not show a clear impact of explanations on perceived workload [2]. In the case of AID systems, the existing workload, contrary to their initial purpose, is partly a major problem that could motivate dropouts. In addition, unreliable integration of sensor technology still contributes to the frequent negative perceived interaction with the system based on alarms [79]. Therefore, when developing explanations or other approaches to increase the traceability of results of intelligent systems, the objective and subjective workload should be controlled.

Additionally, information or explanations can influence trust in intelligent systems [9, 106, 124]. In order for trust to be relevant, risk needs to be present [55]. The incorrect dosing of insulin by an AID system can result in significant health consequences, which is why trust can not only be investigated in the present use case but is also addressed as a prerequisite and challenge for AID use [64]. In this context, clinical reviews, as required from professionals in studies regarding medical AI systems [48], are one way to provide evidence of trustworthiness and thus increase "extrinsic trust" [55]. However, clinical evidence does not affect the traceability of systems. Experienced traceability allows for "intrinsic trust" and, as discussed, the possibility of cooperation. Therefore, human factors research calls for studies on trust in AID systems in dependence of explanations as a suitable means to support intrinsic trust.

Findings in literature on the beneficial effects of explanations are still inconclusive, i.e., different studies observe that the use of explanations did not lead to an objective change in observed behavior. For example, [7] could not find better predictions of AI outcomes even though additional explanations were offered. Similarly, [10] showed that explanations did not significantly increase the joint performance of AI and humans in judging texts. Aggravation of this problem is shown by [29] and [36], where explanations are positioned as "placebic explanations" or even as "dark pattern explanations": these explanations do not contain any information to increase transparency, but induce a better experience of the interaction, e.g., in terms of perceived trustworthiness, adversely leading to "unwarranted trust". This could result in overconfidence and thus an unjustifiably high reliance on, e.g., the AID system. Thus, rather than empowering users, explanations could give them a false sense of security. Especially in the automated delivery of drugs such as insulin, interactions must be designed to prevent the development of overconfidence. Accordingly, the study of objective and subjective measures together in experiments is crucial in the human-centered development of AID systems.

2.3 From Situation Awareness to Subjective Information Processing Awareness

To adequately address human-centered research questions in AID systems, instruments to assess traceability-related facets of user experiences of a system's results are necessary. In recent years, different scales to evaluate XAI have been proposed. [51] gave an overview of user experience metrics for XAI, introducing the Explanation Satisfaction Scale (ESS). The ESS was developed to measure the subjective quality of explanations provided by an intelligent system. Being based on multiple existing methods from the field of trust in automation (such as [56]), it incorporates both affective as well as cognitive implications of explanations (see [75]). The ESS is meant for experts constructing and developing AI systems or experienced users, as they need to rate e.g. the usefulness of results. In iterative development, also a quick interaction with systems needs to provide sufficient data to guide further development. An additional scale allowing inexperienced users, e.g., first time customers and end-users, to participate is crucial for XAI research because usage of AI-based systems is not limited to experts. Another scale addressing system traceability specifically designed for the medical domain is the System Causability Scale (SCS) from [54]. The SCS focuses on a quick overview of the impact of explanations and thus also captures different dimensions, e.g., to what extent users see explanations as transferable to others or whether the explanations fit their own knowledge base. While this allows for a quick general assessment, it is not yet clear to what extent the SCS can also be used for specific, theory-driven questions, e.g., about the traceability of certain decisions. As [125] elaborate in their review, the usability of measurement methods for evaluating explanations depends on the user group, the experimental design, and also the specific properties of the explanation. All in all, existing instruments of XAI research for surveying the subjective effects of XAI often refer directly to the added interaction elements, i.e., explanations given by the system [51, 54].

While these instruments could be used in the selection of appropriate explanations, especially at the beginning of the design process or in formative evaluations, a direct comparison, e.g., to a baseline without explanations may be difficult. To address experimental designs with, e.g., a control group, an instrument that aims to measure the subjective effects of explanations and relates to experienced traceability of automated systems rather than directly evaluate explanations themselves would be advantageous. For this purpose we derive *Subjective Information Processing Awareness* (SIPA) [101] from Situation Awareness Theory. SIPA describes "the experience of being enabled by a system to perceive, understand and predict its information processing" [101]. When users act within a dynamic system, they make situation assessments [38], which result in a user state that has been established as SA. SA Theory postulates three levels within this assessment: 1) perception, where the state of environmental information in the current situation is perceived, 2) understanding, where comprehension of the current situation is formed and 3) projection, where future states of the situation are predicted. Previous work on automation demonstrates how SA may play an important role for XAI research: For example, low SA could be the reason for missing anticipation when information needs to be communicated in order to ensure cooperation [108]. SA loss is a known problem in existing research in human-automation interaction [88]. Hence, understanding the effects of automation on SA is important and applicable to XAI. However, current methods to survey SA have often focused on the interaction's context. On the other hand, SIPA focuses on the transparency of relevant elements, understandability, and predictability of information processing as it is relevant for trust and traceability of AID systems.

While Situation Awareness focuses on processes within the person, the goal of the SIPA scale is to describe the experience of system properties that lead to SIPA. These can be built up analogously to Situation Awareness. Instead of Perception, the first facet of the SIPA scale is experienced transparency, which describes the extent to which the system interaction allows the user to perceive all relevant elements for information processing. Hence, "Understanding" and "Prediction" can analogously be positioned as "experienced understandability" and "experienced predictability". The facets adopted in the SIPA scale are thus grounded in the levels described in SA theory and can be clearly placed within the broad discussion of the definition of, e.g., transparency [26]. Thus, transparency, as defined in the SIPA scale, does not refer to, e.g., goals of the developer or global information on, e.g., training of the model, but to the person's experienced accessibility to information to which the system has access.

To ground the specific items of the SIPA scale in SA theory, we examined different SA scales assessing subjective (c.f. [114]) as well as objective SA (cf. [37]). The items of the scale were developed on the basis of these questionnaires as well as theoretical explanations of situation awareness (as e.g., [39, 123]) and discussed by various experts from the field of engineering psychology. The scale, initially developed with 12 items [101], was shortened by multiple, empirically supported iterations to 6 items. Two of the items are assigned to each of the facets of SIPA. While reverse-coded items were sparingly integrated with the original generation of items, these showed the negative effects discussed in [120]. After weighing the comprehensibility of the scale against the potential negative effects of uniformly one-sided items, no reverse-coded item was included in the 6-item scale - also on the basis of qualitative comments from users.

3 PRESENT RESEARCH

Based on the research issues presented above, hypotheses were derived for the present study. For the hypotheses H1 - H3 the level of information disclosure is the independent variable, while SIPA, the time-on-task and the subjective workload are the dependent variables.

H1: SIPA increases when there is an increase in relevant explaining information disclosed by an intelligent system

H2: Time-on-task increases when there is an increase in relevant explaining information provided by an intelligent System

H3: Subjective workload increases when there is an increase in relevant explaining information provided by an intelligent system

Further, we assume that the dependent variable SIPA increases over time, regardless of the condition, as individuals are given repeated opportunities to make assumptions about the system and correct their mental model.

H4: SIPA increases with increasing observations

As mentioned above, we expect a close relationship between SIPA and trust, since, for example, the experienced predictability of a system as depicted via SIPA is a crucial influencing variable for trust. Furthermore, we expect a strong correlation with ESS due to the similarity of the underlying constructs.

H5a: SIPA and trust correlate moderately to strongly

H5b: SIPA and explanation satisfaction correlate moderately to strongly

Hypotheses H6 - H10 relate to participants' performance on the prediction task or the effects of the prediction task. Here, the prediction of insulin needs calculated by the AID system represents a measurement dependent on the correctness of the participant's mental model. Based on previously discussed theories in the area of cooperation, we hypothesize in H6 - H9 that higher availability of information leads to better SIPA and to better prediction. Additionally, we expected the SIPA value to rise in the performance block.

H6: Higher SIPA ratings before the performance block correlate with better performance in the prediction task

H7: Higher levels of information disclosure lead to better performance in the prediction task

H8: SIPA increases over the course of the performance block

The influence of intra-individual differences (such as attitude towards AI or duration of diabetes) could affect the user experience of an AID system. To assess the inclusiveness of explanations, we formulate the following research question for exploratory analysis:

EQ: How are intra-individual differences related to SIPA ratings and performance in the prediction task?

4 METHOD

We conducted an AID simulation experiment among people living with DMT1. Specifically, we examined how different levels of information disclosure affected the participants' experience of an algorithm calculating insulin needs after repeated interaction with varying levels of information disclosure of the system. The study was pre-registered under https://doi.org/10.17605/OSF.IO/NUJTE at OSF [42]. Changes in the planned and performed analyses are described under Results.

4.1 Participants

80 participants with DMT1 completed the experiment. Ethics approval for this study was granted by the Ethics Committee of the University of Lübeck before the start of the experiment (Tracking number: 21-438). Participants volunteered to participate in the study, and informed consent was required. The experiment was implemented using the Labvanced online experiment platform [41]. Participants were instructed to conduct the study only with appropriate screen size, i.e., at desktop computers, laptops or tablets. We recruited DMT1 patients via mailing lists and social media channels (Twitter, Facebook, Instagram) applying convenience-sampling. Participants were compensated €10 for their time in the study due to the approximated duration of 60 minutes. In addition, the three best performing participants could win €80 each. This additional price was applied in order to put an additional incentive for motivation into performance tasks on top of the general compensation.

To safeguard data quality, we defined two exclusion criteria before the experiment and applied these after study completion: (1) Participants with over-long completion times (>2 *SD*, N = 2 with 412 and 319 minutes in comparison to M = 63 of final sample) were excluded because participants were instructed to complete the experiment in one single continuous session. (2) Participants with very low knowledge of DMT1 management were excluded because the experiment required the most correct understanding of the relationships between the factors influencing blood glucose. To screen for diabetes knowledge, we developed ten items (see Appendix C). To be able to assume sufficient uniform knowledge of diabetes management we defined six correct responses (60% to reach a reliable differentiation from chance) as a cutoff criterion for exclusion prior to the experiment (n = 1 excluded with knowledge score = 4, final sample with M = 7.89 and *SD* = 0.78). In addition to these pre-defined criteria, we observed in the first data inspection that some users reported the same rating for all items in the observation blocks and excluded them to avoid invalid data being part of the analysis. Furthermore, in the prediction task, we observed users to only respond with "0" or positive values in the prediction class, which caused biased results for the prediction. Overall 7 participants were removed based on those additional criteria.

The final sample consisted of 70 participants ranging from 18 to 61 years (M = 28.9, *SD* = 10.5). 49 participants identified themselves as female (70.0% of the sample), 20 as male (28.6% of the sample) and one person as neither. To better classify the sample in relation to the general population with regard to at least one fundamental facet of user diversity (i.e., diversity in human-technology interaction), the Affinity for Technology Interaction scale [43] was assessed. Our sample had a wide range (from 1.22 to 5.67) with an average value of 4.11 being well in the medium range (possible ATI score range = 1-6) yet somewhat higher than reported for the general population (3.5 as described in [43]). Yet, it has to be noted that the average ATI score in the population of AID users is not known (e.g., there is a chance that low-ATI patients are more reluctant to adopting an AID therapy or treatment). The average duration of diabetes was 14 years (*SD* = 10.1, *Range* = 1 - 44) which is similar to distributions of recent clinical studies for AID systems, as for example [12]. Only n = 9 participants stated to have previous knowledge with AID systems. These were evenly distributed across the groups and showed no correlation with performance in the prediction task (all p > .050).

4.2 Experimental Environment

To create an experimental environment we developed an AID simulation system that was designed to meet three criteria: (1) high ecological validity for a good transferability of the results to the practical application of systems, (2) information, that structurally resembles real dynamics in DMT1 treatment with AID systems as well as (3) high experimental control, which allows the systematic manipulation of independent variables and thus enables the research questions to be addressed. Further, the application had to be sufficiently distinct from existing systems, which could otherwise have led to potential confounding based on existing experience and prior knowledge. The AID simulation was created in three steps described in the following sub-sections: 1) the manual creation of valid training data 2) the training of a basic machine learning model for use in the context of a run-time capable AID simulation, and 3) the generation of static scenarios for a controlled experiment.

4.2.1 Development of Artificial Training Data for AID simulation. An artificial data set of information relevant for AID systems was developed to be independent of individual medical data and the complications that come with it in terms of using personal health data. Each instance consisted of 12 different attributes and the insulin requirement. The individual data sets represent different individuals and therefore contain individualized factors as attributes, such as the amount of correction for excessive glucose levels. All attributes and their meaning are found in Appendix A. Negative insulin needs refer to the need to take in carbohydrates when, e.g., too much insulin is in the body. The different attributes are based on data that is already used in various clinically tested AID systems [12, 80]. After creation, the data set was reviewed by two independent diabetologists. Both independently rated the data set as plausible. In total, over 480 instances were created, with 400 to train and test a model.

The attributes have been divided into three different groups, following the approach discussed in Related Work: (1) information provided to the system by the user depending on the situation or automatically determined by the system and **representing physiological variables** influencing the amount of insulin, (2) information representing general or dynamic therapy **goals or preferences of the user**, and (3) **information learned by the algorithm, which provides information about the calculated insulin sensitivity** and thus factors influencing the outcome of the AID system. The information of the first group is oriented to give one (1) common ground about information that both human and machine absolutely need for cooperative action. The information of the second group shows which possibilities the system has for (2) implementing user preferences and can thus give users information about the extent of directability. While all information increases the predictability of the system, the information from the third group represents influencing factors for the concrete (3) computation of the system.

4.2.2 Training of random forest model for AID simulation. Subsequently, a model was trained based on the data. To predict insulin needs based on the dedicated attributes as input parameters, a random forest regressor was implemented [103], see also [84]. A train-test-split where 25% or the data was reserved for testing was used, resulting in 4 datasets: X_train, X_test, y_train, y_test. The X datasets include the input parameters for the regressor, while the y datasets only contain the corresponding target values (results).

Through a grid-search cross validation algorithm, a (on average) best set of hyperparameters for the random forest were found to be: 80 estimators and 10 max depth. These parameters are used for the construction of the random forest and control the number of trees in the forest and the max depth of those trees. A lower number of trees would have resulted in an underfitted model, while a higher number of trees (> 100) would not have increased performance further.

Table 1. Overview of attributes used in the simulation

	Attributes
Low Information Disclosure (LowID)	Current Tissue Glucose Current Insulin in Body Current Carbohydrates in Body Current Activity
Medium Information Disclosure (MedID)	Tissue Glucose Target Avoid Hypoglycemia Duration of Insulin Effect Correction Intensity
High Information Disclosure (HighID)	Risk of Hypoglycemia in next hour Blood Glucose lowering per 1 Unit Insulin Insulin Units per 10 grams Carbohydrates Predicted Exercise

Table 2. Hyperparameters of applied random forest model

Mean Absolute Error (MAE)	3.0250
Mean Squared Error (MSE)	13.6905
Root Mean Squared Error (RMSE)	3.7001
Mean Absolute Percentage Error (MAPE)	1.5254
Explained Variance Score	0.3922
Max Error	7.9771
Median Absolute Error	2.2023
R^2	0.3887

The maximal tree depth of 10 shows a good performance for the dataset at hand, while deeper trees are more prone to noise in the data.

The random forest was then fitted to the training data sets (X,y) with the hyperparameters. The regression model exhibits metrics when comparing predicted values with real result values (y_pred, y_test) as shown in 2.

4.2.3 Generation of scenarios for a simulation-based experiment. The AID simulation was used to generate scenarios for an experiment. The interactive input of individual data was excluded for this experiment in order to 1) have uniform scenarios for each participant and thus avoid biases due to different inputs 2) to focus on scenarios close to the application and 3) to reduce the risk of technical problems in the ongoing experiment in the context of the experiment conducted online.

To create scenarios, calculated insulin needs were removed from the 80 remaining instances of the previously described data set and used as inputs for the AID simulation. The outputs were saved as screenshots, with all 80 scenarios saved in three different formats and used in the experiment as conditions: (1) low information disclosure (LowID), (2) medium information disclosure (MedID), or (3) high information disclosure (HighID). The allocation of information is based on the groups described above and is presented in 1.

The resulting interfaces can be seen in Figure 1. Participants consistently saw only one of these conditions throughout the experiment, in both the observation and performance blocks. Because of feedback in pre-tests, the concept of correction strength was explained to all participants from MedID and HighID before each block of stimuli.

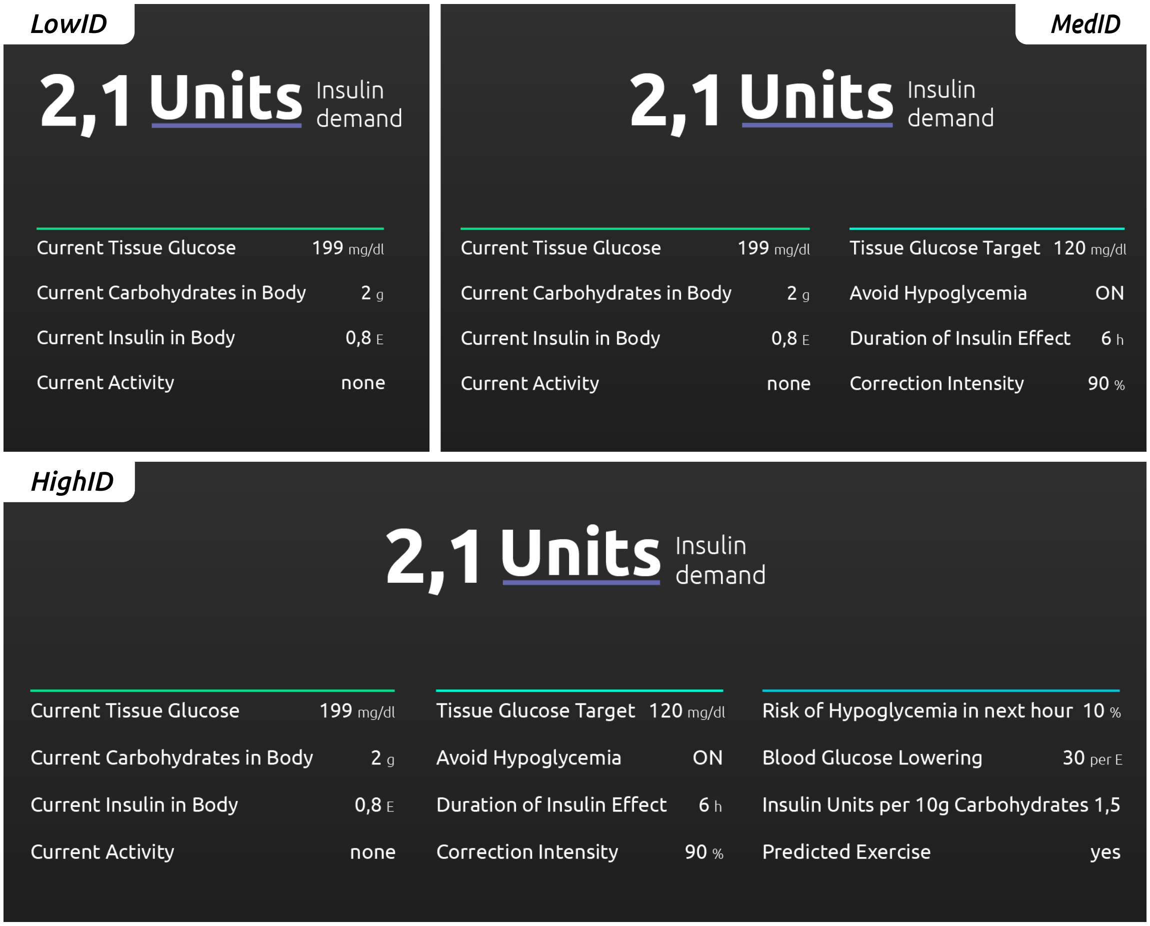

Fig. 1. Stimuli from the study as they were shown to participants for the three conditions: LowID, MedID and HighID.

4.3 Measures

4.3.1 SIPA Scale. The SIPA scale as a measure to assess users' experience while interacting with intelligent systems was used to examine effects of different levels of information disclosure. The goal for the development of the SIPA scale was to construct a highly economical scale closely linked to SA but focused on an application in intelligent automation, respectively XAI. Additionally, the scale is specifically designed to assess the 3 facets of SIPA as described above (see Related Work) with two items for each facet (1 and 2 for transparency, 3 and 4 for understandably and 5 and 6 for predictability). All items are shown in Table 3.

The 6-item SIPA scale uses a 6-point Likert response scale from completely disagree = 1, largely disagree = 2, slightly disagree = 3, slightly agree = 4, largely agree = 5, to completely agree = 6. The SIPA scale introduced in the present paper was additionally tested over all points of measurement of SIPA for three-factor structure to examine if separate evaluation of the 3 individual facets of SIPA was supported. Here, the approach to analyze 3 facets received support based on a confirmatory factor analysis demonstrating a good fit with χ^2 (6) = 7.49, p = .278, *CFI* = .997, *TLI* = .992,

Table 3. 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
01	It was transparent to me which information was collected by the system.						
02	The information that the system could acquire was observable for me.						
03	It was understandable to me how the collected information led to the result.						
04	The system's information processing was comprehensible to me.						
05	With the information accessible for me, the results was foreseeable for me.						
06	The system's information processing was predictable for me.						

RMSE = .06 (90% CI: .00, .17). The correlation between transparency and understandability was significant (r_S = .64, $p <$ .001), which was also true for the correlation between transparency and predictability (r_S = .53, $p <$ .001) as well as for the correlation between understandability and predictability (r_S = .79, $p <$ .001).

4.3.2 User diversity variables. User diversity can have a significant impact on the individual user experience and, for example, influence initial trust in a system [8]. To examine the role of user diversity on the experience of interaction with an AID system, two additional variables were collected: 1) affinity for technology interaction (ATI) [43], which is based on the personality trait need for cognition [24] and describes the individual tendency to actively engage in intensive technology interaction. ATI was measured with a scale validated in various large samples [43], and the present sample was assessed as rather affine to interact with technology (see section participants above). Furthermore, the individual attitude towards artificial intelligence was surveyed. To this end, a brief definition of artificial intelligence was first given (see Appendix). Based on this, six statements from the Internet Attitude Scale [59] were adapted, with "Internet" as the subject being replaced by "Artificial Intelligence" in all used questions (see Table X). A mean value was calculated to evaluate the Artificial Intelligence Attitude (AIA). In addition, questions on prior diabetes knowledge were used (see Appendix). This included 10 different statements about the treatment of diabetes to ensure that the results of the study were not affected by significant differences in prior knowledge about the treatment of diabetes. Everyday examples of the treatment of type 1 diabetes or questions about how insulin works were used. Finally, the duration of diabetes in years was requested.

4.3.3 Subjective Measures for trust, satisfaction & workload. In addition to the SIPA scale, subjective variables were collected with economical scales. The Facets of System Trustworthiness Scale (FOST) [116] was used to measure trust. With 5 items, this can be used much more economically in a repeated-measures experiment compared to, for example, the more widely used scale of [56]. As for trust, the mean value of the FOST items was calculated for each point of measurement.

The perceived workload was collected through the NASA Task-Load-Index (NASA-TLX) [49]. However, due to the experimental conditions, not all dimensions of the NASA-TLX were used, but the question about perceived physical workload was excluded. Furthermore, the results for effort, mental demand, and time demand were summed to a mean value. Experienced frustration was evaluated independently of other values. The estimation of own performance was only used as a confidence measure after the subjects themselves made a prediction of the algorithm's results. Additionally to SIPA and trust, the Explanation Satisfaction Scale (ESS) was measured to allow a comparison to another scale examining the quality of explanations [51]. The ESS was developed to measure the subjective quality of explanations provided by an intelligent system.

4.3.4 Objectives Measures for Performance & Time-On-Task. In the present experiment, time-on-task (TOT) and a performance indicator were assessed as objective variables. For TOT, the time that the users spent in the different task blocks was measured in seconds. For the analysis, the sum of the time in seconds was calculated. For the assessment of the performance, 20 of the 80 stimuli created with the AID simulation environment were changed in such a way that no prediction of the algorithm was displayed, but the different levels of information disclosure (depending on the condition). Participants were prompted to estimate the output of the algorithm (this could be negative or positive with one decimal place, or the "0"). The deviation of each estimate was determined per person and a mean value was calculated, which was used as an indicator of performance.

4.4 Procedure

The study was conducted in German. In the beginning, the participants were instructed to watch a video where an instructor to the study explained the purpose of the study as well as the tasks. The spoken text was displayed later in written form and could be read again if needed. Afterwards informed consent was obtained from all participants. The experiment was conducted in multiple segments as depicted in 2: first, demographic data was collected (1); then, knowledge questions about diabetes were asked to minimize effects of divergent prior knowledge (2). Subsequently, all participants were randomly assigned to one of three conditions - low, medium, or high level of information disclosure. Depending on this, 15 stimuli were shown in random order in an (3) Observation Block, after which SIPA, FOST, and the NASA-TLX were queried. Three additional observation blocks with other stimuli followed by SIPA, FOST, and NASA-TLX followed (blocks 4-6). Subsequently, the ESS was surveyed (7). Finally, in a performance block (8), 20 stimuli were presented in which participants had to estimate for themselves the insulin needs calculated by the algorithm. The stimuli again differed in the level of information disclosure and were stimuli the participants did not see before. However, the same instances were shown to all participants in a randomized order (i.e., each participant saw the same tasks, but with different information being presented and in different sequence depending on the condition they were assigned to). SIPA, FOST, and NASA-TLX were then collected again. Furthermore, the time for each observation block as well as for the performance block was collected. Depending on the individual deviation from the correct calculated insulin needs, a code was created and displayed to the participants in the last frame of the study. To ensure the anonymity of all subjects the code only corresponded with the deviation and didn't give any indication to personal information.

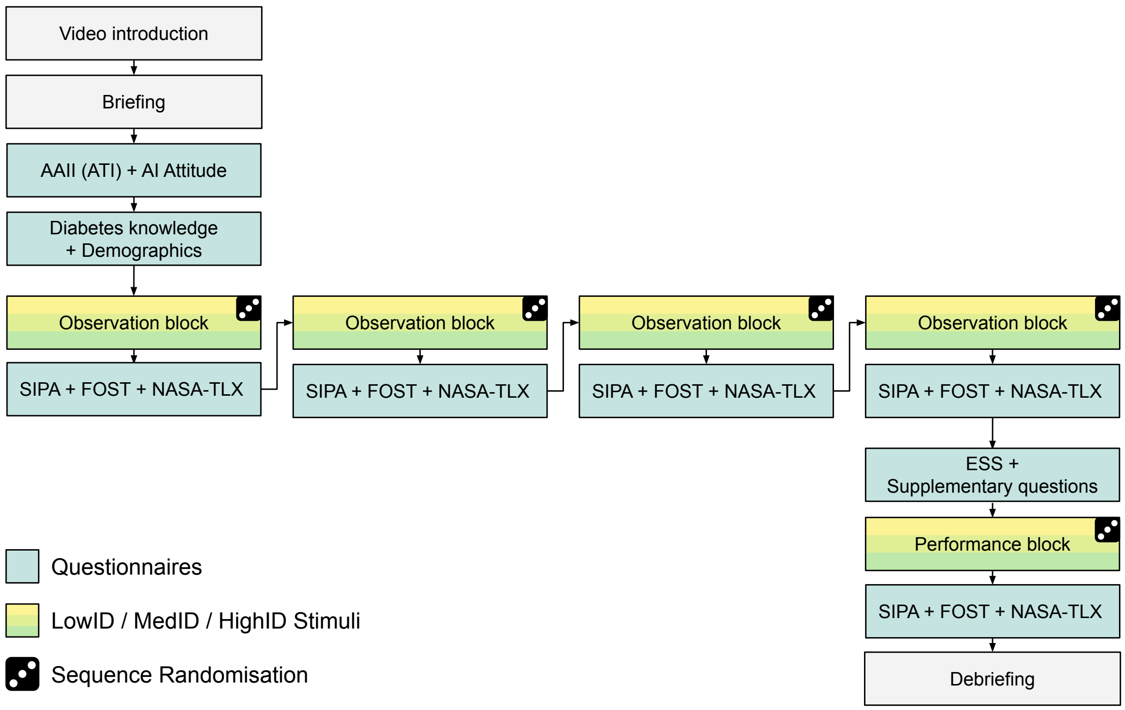

Fig. 2. Overview of course of the experiment.

5 RESULTS

As a direct test of our hypotheses we applied contrast analysis, which allows for a more precise testing of hypotheses [22, 122]. However, that approach was different to our pre-registration where we only planned to conduct a less precise omnibus testing (i.e., ANOVAs), yet omnibus F-test are inefficient in order to extract the effects the present study aims to examine. The core hypotheses H1 - H5 related to the development of user experience in repeated observations were part of the pre-registration. Additional Hypotheses H6 - H10 relate to performance or self-assessment of performance and were not pre-registered. One-tailed t-tests were conducted to assess the hypotheses. All p-values were corrected for family-wise error [13] for each hypothesis and variable using the Bonferroni-Holm correction [53]. Despite random assignment, not all groups are exactly equally distributed (n = 24 for LowID, n = 22 for MedID, and n = 24 for HighID). Since multiple variables studied were not normally distributed (or no linearity could be assumed), Spearman's Rho was calculated for all correlations and interpreted accordingly depicted as r_S. Effect sizes for r and r_S were interpreted based on [44, 97], effect sizes for d were analyzed according to [30] with respect to [44]. Cohen's d was reported for contrast analysis of dependent measures instead of Hedge's g, because both are almost equal in sample sizes greater than 20 [62].

5.1 H1: SIPA increases when there is an increase in relevant explaining information disclosed by an intelligent system

H1 was examined using multiple contrast analyses [22, 122], one for each SIPA facet (transparency, understandability, and predictability) and for each point of measurement. The different amounts of information disclosed to each group and the corresponding relationship between attributes was used to determine the weights (i.e., lambda values). It is assumed that each attribute (i.e., a total of LowID: 4, MedID: 8, or HighID: 12) can be related to each other attribute seen in one condition. The number of relations between attributes is given by the binomial coefficient (i.e., number of attributes over two). Thus, the number of relations between attributes is for LowID = 6, for MedID = 28, and for HighID = 66. Following [22] to calculate the weights, the following lambda values for the contrast analysis were defined: $\lambda_{LowID} = -2.5$, $\lambda_{MedID} = -0.5$, $\lambda_{HighID} = 3$. Table 4 shows the *t*-statistics, the corrected *p*-value as well as $r_{(effect\ size)}$. *M* and *SE* are depicted in Figure 3. All descriptive data can be found in Appendix B. Results regarding the SIPA facet of transparency supported H1 for observation blocks 3-4 and the performance block, while the first observation blocks 1-2 did not show significant effects supporting H1 (see Table 4). The other two SIPA facets understandability and predictability showed weak effects in the expected direction which where all non-significant (except ratings for SIPA understandability after Observation Block 1 and Observation Block 2, which were small but contrary to the hypothesis). Hence, H1 was supported for experienced transparency after considerable experience of the system, yet not directly after the first interaction and not for the more complex systems properties measured by SIPA (i.e., understandability and predictability).

Table 4. H1: Contrast Analyses for each SIPA facet comparing ratings between conditions (LowID, MedID & HighID) for all blocks

	SIPA transparency			SIPA understandability			SIPA predictability		
	t	*p*	$r_{(effect\ size)}$	*t*	*p*	$r_{(effect\ size)}$	*t*	*p*	$r_{(effect\ size)}$
Observation Block 1	0.32	.375	.04	-1.03	.612	-.13	1.14	.258	.14
Observation Block 2	1.89	.063	.23	-0.39	.349	-.05	1.35	.363	.16
Observation Block 3	2.37	.031*	.29	0.64	.786	.08	0.36	.360	.04
Observation Block 4	2.47	.032*	.30	0.56	.578	.07	1.16	.375	.15
Performance Block	2.46	.040*	.29	1.56	.309	.19	2.08	.104	.25

Note. df = 67 for all analyses. * $p < .050$, ** $p < .010$, *** $p < .001$

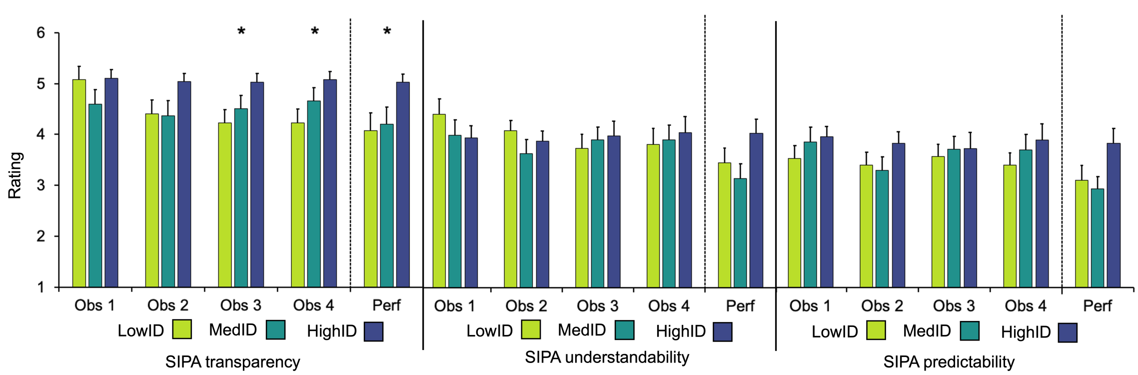

Fig. 3. H1 & H4: Ratings of the SIPA scale for all points of measurement. Bars depict *M* and *SE* for all SIPA facets at each time measured. * indicate $p < .050$ for contrast analysis, as shown in Table 4.

5.2 H2: Time-on-task increases when there is an increase in relevant explaining information provided by an intelligent System

To test H2, multiple contrast analyses were used. The corresponding results can be found in Table 5. Contrary to the hypothesis, there was no significant difference between the groups for all blocks, apart from one exception (performance block). Interestingly, a medium effect aligned with the hypothesis was present in the performance block. Thus, the performance block clearly stands out and supports the hypothesis, while the data of the observation blocks do not.

Table 5. H2: Contrast Analyses comparing time-on-task between conditions (LowID, MedID & HighID) for all blocks

	Time on Task		
	t	p	$r_{(effect\ size)}$
Observation Block 1	1.83	.107	.22
Observation Block 2	1.03	.153	.13
Observation Block 3	1.51	.136	.19
Observation Block 4	1.82	.146	.22
Performance Block	4.20	$< .001^{***}$	.47

Note. * p <.050, ** p <.010, *** p<.001

5.3 H3: Subjective workload increases when there is an increase in relevant explaining information provided by an intelligent system

To test H3, multiple contrast analyses were used. The corresponding results can be found in Table 6. Contrary to the hypothesis, in all blocks workload ratings were not significantly higher in conditions with more information. Indeed, negative signs in t-statistics at all points of measurement indicate, that the effect was actually in the other direction (i.e., more information disclosure decreases workload). In fact, an exploratory re-calculation of the contrast with inverted weights (i.e., $\lambda_{\text{LowID}} = 3$, $\lambda_{\text{MedID}} = -0.5$, $\lambda_{\text{HighID}} = -2.5$.) of the effect would support an oppositely formulated hypothesis, e.g., with $p < .001$ and $r_{(effect\ size)} = .36$ for Observation block 1.

Table 6. H3: Contrast Analyses comparing subjective workload between conditions (LowID, MedID & HighID) for all blocks

	NASA-TLX		
	t	p	$r_{(effect\ size)}$
Observation Block 1	-0.92	.540	.03
Observation Block 2	-1.45	.304	.10
Observation Block 3	-0.69	.492	.04
Observation Block 4	-0.18	.429	.03
Performance Block	-1.64	.264	.12

5.4 H4: SIPA increases with increasing observations

To test H4 multiple contrast analyses were conducted for each SIPA facet (transparency, understandability, and predictability), but followed the contrast analysis for dependent measures [102]. The following weights were used for each analysis: $\lambda_{\text{Observation 1}} = -1.5$, $\lambda_{\text{Observation 2}} = -0.5$, $\lambda_{\text{Observation 3}} = 0.5$ and $\lambda_{\text{Observation 4}} = 1.5$. Table 7 shows the t-statistics, the corrected p-value as well as d. Counter to our hypotheses, SIPA ratings did not increase but decreased and the actual effect of repeated observations was opposite to what we hypothesized. In fact, a follow-up calculation with inverted contrasts significantly supported he assumption of decreasing ratings for transparency with p = .44, while p > .050 for understandability and predictability.

Table 7. H4: Contrast Analyses comparing repeated SIPA ratings for Observation Blocks 1 - 4

	Contrast Analysis for Obs 1-4		
	t	p	d
SIPA transparency	-1.73	.956	0.21
SIPA understandability	-0.95	.827	0.12
SIPA predictability	-0.40	.827	0.05

Note. * p <.050, ** p <.010, *** p<.001

5.5 H5a: SIPA and trust correlate moderately to strongly

To test H5a, the correlation between the FOST scale scores and each SIPA facet was calculated for each point of measurement. The results are shown in Table 8. The range of effect sizes of the correlation across all facets is between r_S = .58 and r_S = .85, which indicates a strong relationship. Overall, the hypothesis can therefore be supported by the data.

Table 8. H5a: Correlations between trust and SIPA facets for each point of measurement

		SIPA					
		Transparency		Understandability		Predictability	
		r_S	p	r_S	p	r_S	p
Trust	Observation Block 1	.58	< .001***	.76	< .001***	.64	< .001***
	Observation Block 2	.60	< .001***	.85	< .001***	.80	< .001***
	Observation Block 3	.64	< .001***	.84	< .001***	.82	< .001***
	Observation Block 4	.65	< .001***	.84	< .001***	.79	< .001***
	Performance Block	.72	< .001***	.81	< .001***	.76	< .001***

Note. * p <.050, ** p <.010, *** p<.001

5.6 H5b: SIPA and explanation satisfaction correlate moderately to strongly

To test H5b, the correlation calculated between each SIPA facet for Observation Block 4 with ESS was calculated. All facets of SIPA showed a significant correlation (all p < .001), with transparency r_S = .57, understandability r_S = .67 and predictability with r_S = .65 indicating a strong correlation, which supports the hypothesis.

5.7 H6: Higher SIPA ratings before the performance block correlate with better performance in the prediction task

To test H6, correlation between each SIPA facet for Observation Block 4 with the overall performance was calculated. No significant correlation was found for transparency (r_S = -.11, p = .850), understandability (r_S = -.17, p = .355) or predictability (r_S = -.08, p = .731). Thus, a correlation between the SIPA ratings before the performance block and the performance cannot be assumed and the hypothesis is not supported.

5.8 H7: Higher levels of information disclosure lead to better performance in the prediction task

To test H7 a contrast analysis was performed. The weights correspond to the weights used in H1 with λ_{LowID} = -2.5, λ_{MedID} = -0.5, λ_{HighID} = 3. A one-tailed significance test with ($t(67) = 1.21$, $p = .116$, $r_{(effect\ size)} = .15$) did not detect a significant difference between the groups, thus there was no support for the hypothesis.

5.9 H8: SIPA increases over the course of the performance block

To test H8, multiple contrast analyses were conducted for each SIPA facet following the contrast analysis for depended measures. The following weights were used for each analysis: $\lambda_{Observation\ 4}$ = -1.5, and $\lambda_{Performance}$ = 1.5. A one-sample t-test against zero was performed for all contrasts. Table 9 shows the t-statistics, the corrected p-value as well as d.

The hypothesis is not supported by the results for any of the SIPA facets. However, all facets show a high negative t-statistic, which suggests that the contrast was chosen in opposite to the real data. This corresponds to the descriptive observation that there was not a successive increase but decrease in the SIPA ratings for all facets. The calculated effect sizes also indicate a relevant effect at the boundary between small and medium effect. Under the assumption of opposite contrasts, significant effects are shown for transparency (p = .014), understandability (p = .010) and also predictability (p = .016).

Table 9. H8: Contrast Analysis comparing SIPA facets before and after performance block

	Contrast Analysis for Obs 1-4		
	t	*p*	*d*
SIPA transparency	-2.26	.986	- 0.28
SIPA understandability	-2.40	.991	- 0.29
SIPA predictability	-2.19	.984	- 0.27

Note. * p <.050, ** p <.010, *** p <.001

5.10 EQ: Explorative Analysis of Individual Differences

To examine the relation between individual differences in human-AI cooperation and user experience, correlations between person characteristics (ATI, AIA, duration of diabetes) and SIPA ratings as well as performance were calculated. The measurements for Observation Block 1 and the performance block were analyzed in order to keep the number of tests (and the resulting loss of power due to correction) low. All values are shown in Table 10. There was no correlation between the duration of the disease and the SIPA ratings or the performance. With regard to the ATI values, no correlation can be found at the beginning of the experiment. At the last time point, there is a small to moderate effect (for SIPA transparency and SIPA predictability). For AIA, no significant effects are found at the end of the study, but at the beginning of the experiment there are moderate, significant correlations with SIPA transparency and SIPA understandability. Neither ATI nor AIA show a significant relationship with performance.

6 DISCUSSION

6.1 Summary of Results

The objective of the present research was to examine the effects of explanations that differ in the amount of disclosed information as well as the effect of repeated interaction on users' subjective perception of trust and traceability in AID

Table 10. Results of Explorative Analysis

			ATI		AIA		Duration of diabetes	
			r_S	p	r_S	p	r_S	p
SIPA	Transparency	Observation Block 1	.29	.080	.38	.024*	-.29	.098
		Performance Block	.42	.007**	.29	.144	-.10	> .999
	Understandability	Observation Block 1	.24	.150	.36	.115	-.25	.240
		Performance Block	.24	.192	.14	.256	-.01	.961
	Predictability	Observation Block 1	.23	.104	.27	.014*	-.21	.410
		Performance Block	.35	.018*	.18	.099	.03	> .999
Performance			-.04	.731	.05	.666	-.07	> .999

Note. * p <.050, ** p <.010, *** p <.001

systems. Contrast analyses were performed to test directional hypotheses related to the dependent variables SIPA, TOT, and subjective workload.

While results showed a weak tendency for users in the HighID condition to report higher SIPA ratings than users in the LowID condition, the assumed contrast (increasing SIPA ratings with increasing quantity of disclosed information) was only significant for SIPA transparency after multiple interactions (i.e., after 45 observations) and aligned with hypothesis **H1**. The time users spent on the prediction task was more than twice as high for users in the HighID condition then for users in the LowID condition. Thus, a significant raise of TOT based on higher information disclosure as stated in **(H2)** could be found when participants were asked to predict the insulin needs calculated by the system. In contrast, only non-significant and slight differences were found when people were instructed to observe stimuli displaying the insulin needs calculation. Although the subjective workload did not increase significantly with the level of information disclosure as assumed **(H3)**, an unexpected effect emerged: the perceived workload was higher for the LowID condition than for the HighID, in some cases more than one standard deviation higher. The development of the SIPA rating over time also shows, contrary to our expectation **(H4)**, a decrease This effect was small for SIPA transparency, while only negligible effects can be observed in the other facets. A strong correlation between all SIPA facets and trust **(H5a)** as well as between all SIPA facets and explanation satisfaction **(H5b)** indicates high convergent validity for the SIPA scale. SIPA ratings prior to the performance block did not correlate with performance itself and also showed very small effects **(H6)**, although SIPA transparency ratings differed significantly before observation for different levels of information disclosure. Although more information were available in the MedId and HighID than in the LowID condition, participants in the MedID or HighID condition did not perform significantly better than participants in the LowID condition **(H7)**. The prediction task in the performance block did not lead to an increase in SIPA, but resulted in lower SIPA scores in all facets with a medium to strong effect size **(H8)**. Analysis of intra-individual correlations with SIPA revealed that SIPA was significantly related to attitudes toward AI after the Observation Block1, while ATI showed a significant influence after the Performance Block **(EQ)**.

6.2 Effects of information disclosure on user experience and cooperation in AID systems

One focus of the present work was to investigate the effect of different levels of information disclosure on the user experience of AID systems. However, higher information disclosure did not affect SIPA immediately but led to a significant difference in perceived transparency only after 45 observations. The delayed decrease in SIPA transparency ratings suggests that a valid measurement of subjective variables may need an experimental design with sufficient repetitions (cf. as for trust [46, 52]). While individuals in the LowID condition started with a comparably high level of

SIPA transparency, the observed decrease could be related, for example, to the fact that only repeated observations allowed them to recognize that not all necessary information was available. [98] describe that a person's mental model is used to form expectations about the outcome, e.g., of cooperation with automation. As for trust (cf. [50]), individual differences could affect the initial SIPA rating, and only measurement after a system-dependent number of interactions can reveal differences between systems. This is also exemplified by the co-relationship between AIA and SIPA transparency at initial observation and after performance, which indicates that explanations may be able to offset the effects of initial mistrust of individuals. The relationship between attitudes toward AI systems (such as AIA) and other user diversity factors (such as education level or access to technology) represents another research challenge to explore the effects of explanations more in depth.

Another reason, why participants in HighID or MedID did not show better prediction performance, could be information overload. Information overload occurs when an increase in the available amount of information leads to negative results, e.g. a decrease of performance or subjective consequences (e.g. as experienced cognitive demand or stress) for the user [70]. Although there was three times as much information available in the HighID condition than in the LowID condition, the TOT for the observation blocks did not differ significantly between the groups. [5] assume that a high information workload can lead to the use of heuristics (e.g. the representativeness heuristic) or increase the probability of users to make biased decisions. This effect is opposed to one goal of XAI design, which is to mitigate errors based on heuristic decision making [118]. In our AID simulation experiment, the use of heuristics while observing might have been higher for the HighID condition than for the LowID condition. This could explain why TOT did not increase (for the observation blocks) though more attributes were presented. The results of the NASA-TLX on subjective workload allow a parallel conclusion: experienced time demand, cognitive demand, and effort showed no difference between the conditions. It is very unlikely that the participants of the HighID condition did not notice or ignored the additional information, as they partly referred to it in the qualitative comments. While being already discussed [93, 118], the extent to which explanations or the additional information available through explanations create an information overload and thus influence, for example, the use of heuristics in the evaluation (see also [35]) of an AID system still needs to be investigated more clearly and for users of different levels of expertise. [113] found that, for example, the expertise of users can decrease the probability that they will use heuristics. However, AID systems, in particular, have great potential for individuals with problematic long-term metrics, which in turn may often be due to low engagement with and care for the disease. For an inclusive design of AID systems, effects of explanations for less experienced users must be understood and avoided, in case they cause, e.g., limited transparency. Representations that lead to a heuristic assessment due to information overload could thus encounter users for whom a heuristic assessment could be particularly problematic. All in all, when designing XAI and in order to act responsible, developers should consider, that more access to information may be harmful to transparency and elaborated context analyses are needed to understand how users will interpret and utilize information or explanations.

Finally, the qualitative results point to another problem, as participants from the LowID conditions explicitly ask for information that was presented to the other groups, e.g., LowID-1: "Please add probability of hypoglycemia or intensity of correction" or LowID-2: "Please show correction quotas for glucose and carbohydrates". However, the results of the experiment suggest that this does not necessarily allow for higher SIPA or better prediction. In order to achieve higher SIPA, the individual pieces of information presumably need to be put into better proportion, as HighID-1 expresses: "I need refined information, how much insulin is given to correct glucose levels and how much is given for food". The requirement for a more mathematical description could be due to the fact that users apply their mental models of how they would solve the problem without an AID system to the system's information processing. In the field of AID systems, users potentially perform a complex calculation, through which they have certain expectations, as HighID-2 states: "I would like to see highlighting of factors that are particularly decisive for the calculation at that moment". In future explanations of AID systems, the representation of the calculation should be as close as possible to the calculation performed by the users (as depicted by [118]) in order to empower users to assess the system's information processing. This would also meet a central criterion for cooperation, where an adequate communication of information requires partners to anticipate the relevance of the information for the task of the cooperating partner.

6.3 Fit of performance measures and subjective measures in XAI

In our experiment, the participants' own assessment of the system's traceability does not correlate with their ability to predict the system's calculation. This is a worrisome correlation, since in the best case false expectations arise and people lose confidence in the system. A more serious consequence could be, for example, a misjudgment of the system's performance in extreme situations and the development of overconfidence. Several studies [77, 78] on trust in automation show that a lack of calibration between subjective ratings and objective scores is a well-known phenomenon. This miscalibration can lead to significant problems, e.g., complacency arises and thus the users attribute more competencies to the system than it possesses [88] - which is described as an abuse of the system [87]. On the other hand, mistrust can lead to a misuse of the system [87] - in the case of the AID system, suggestions of the system could be corrected frequently and thus lead to an increase of the workload instead of a reduction. Both forms of lack of calibration are significant problems in the AID domain and could help to explain dropout rates [79]. The calibration of SIPA and the correct prediction of an outcome is theoretically more direct than the calibration between prediction and trust (e.g., I can trust the technical competence of a system without understanding how it works, see [75]) and can be used in future studies to show the miscalibration between user experience and the correctness of one's mental model. [98] describe a user's mental model as a 'mechanism whereby humans generate descriptions of system purpose and form, explanations of system functioning and observed system states, and predictions of future system states'. This is also in line with central concepts of SA theory or the idea of so-called situation models [11]: here, mental changes are carried out in order to assess the effects of one's own actions. However, figuring out how changing input variables affect the outcome of an AIs information processing, may be complicated in the case of static explanations (c.f. [1]). Also [27] show that static explanations have a smaller influence on the ability to understand a system than interactive explanations. The latter allows users to build hypotheses on their own and test them, which is the central approach for knowledge acquisition (c.f. [90]). Interactive explanations should therefore be made possible for AID systems (and in other intelligent systems). At the same time, future experiments should focus on observing the formation of hypotheses and their evaluation in the interaction between humans and AI, e.g., to identify when explanations favor confirmation bias or disadvantage individuals with less prior knowledge and how those effects can be mitigated.

This is also supported by the fact that the prediction task had a clear influence on SIPA ratings - all facets of SIPA were reduced, while this was not the case for SIPA understandability and SIPA predictability even after 60 previous repeated (passive) observations. The information provided (i.e., the attributes) was not changed for the performance block. In further studies or development of AID systems, active prediction of AID results should therefore be part of the experimental condition and based thereon considered in training. The role of feedback for SIPA as well as trust should again be considered separately. For example, the diagnosticity [16] or the diagnostic value [121] of certain attributes (i.e., what informativeness they had in determining insulin needs calculated by the system) might have been misjudged by individuals. This could be corrected by feedback or an interactive simulation.

Another obstacle, however, is the information overload discussed above, which could also arise in an interactive simulation. While, e.g., explanations on the basis of "counterfactuals" [81] may be well suited for testing hypotheses, more research needs to examine how larger numbers of, e.g., setting possibilities affect the interaction. In the exemplary case of generative visual models, the cognitive load of the user increases with the number of adjustable settings - without a significant effect on performance [31]. Furthermore, it must be considered whether and which additional information is displayed e.g., in a training context or in a daily use context since these may differ considerably with respect to the available time and cognitive resources. Here, explanations need to be designed for diverse users (i.e., the trainer, which are often medical professionals as well as the patients). The fact that more attributes lead to a higher time requirement for the derivation of a prediction was also shown in the present experiment (see H4, Performance Block). Overall, context-specific prioritization of information must be made, which could be done based on the following questions: 1) Does the representation of attributes/relationships fit the existing mental model of the users? 2) Does the presentation of attributes / contexts allow for hypothesis generation and testing?

6.4 Research & Design Implications for AID systems

For the research of experienced traceability of intelligent systems, the SIPA scale with its facets allows for two central observations: 1) a sufficient number of repeated interactions as well as 2) a differentiation of active interaction from passive observation of explanatory information disclosure are necessary to discuss human-centered AI. The SIPA scale is an appropriate instrument for this context for the following reasons: the SIPA scale shows good scale metrics (i.e., range, standard deviation) on all facets. Additionally, due to high correlation between all 3 SIPA facets also a unidimensional application is possible. Furthermore, the SIPA scale shows a very high convergent validity with measures of perceived trustworthiness and satisfaction with explanations. However, there is a small to medium correlation between ATI and SIPA, and the ATI mean of the present sample is higher than the estimated population mean. Hence, the use of the SIPA scale in groups with lower ATI scores might be different, e.g., shows other correlations with satisfaction. Overall, the SIPA scale with its facets represents a new tool for researching experienced traceability, which can help to underscore and evaluate the effects of explanations on users in detail.

The boundary between Situation Awareness and Performance (i.e., Prediction) has already been raised repeatedly in the discussion of Situation Awareness [89]. While a theoretical discussion of these concepts is beyond the scope of this paper (c.f. [76]), a very high crenelation between SIPA understandability and SIPA predictability suggests that the difference between Understanding and Predicting might be too small to provide an impactful analysis. Studies using other explanatory approaches would need to investigate whether this difference can be amplified. In addition, qualitative comments from users suggest that another facet of Traceability may be relevant - the assessment of the relevance of attributes to the information processing, explicated, e.g., from MedID-2: "Display to what extent which information contributed to the result", which possibly refers to the individual attribute's influence or relevance for the prediction (i.e., diagnosticity, c.f [16]). The extent to which the presented information has a high, subjective diagnosticity could be distinguished from predictability as a facet. For example, an AID system's user might know that providing information about exercise intensity is more important than providing information about the duration of the physical activity. The user would feel able to instruct the AID system to achieve a more precise prediction, regardless of the user's ability to specify the concrete outcome. Especially for the communicative processes in the field of human-AI cooperation, such an additional facet could enable, e.g., what [28] describe as collegiality.

When designing AID systems, the effects on the experienced traceability as well as on workload and performance must be taken into account. The sole disclosure of additional information cannot be seen as a suitable method to improve the user experience or the basis for human-AI cooperation in AID systems. In the given scenarios, the information used from the AID simulation was relevant for the calculation of the system and mimics information that users themselves need for a calculation. The fact that this approach did not offer a significant advantage for the participants of the HighID condition shows how much human-centered research is still necessary in the XAI area. In XAI research explanatory approaches partly refer to the confidence of the model [10, 17, 85] for a certain result or even to meta-information about the model [74]. Depending on their task, such information might have only low significance for the users. This could lead to erroneous conclusions in the future, especially if the methods to evaluate the performance of human-AI cooperation are based on different processes than the processes supported by the explanation. Regardless of how helpful certain methodologies are to AI method developers, users as well as the constructs and requirements relevant to them may be entirely different and need different explanations. Even among the users of a system (in the broadest sense), there might be differences. That is, the information presented in our experiment might help individuals with medical training who, for example, match the model's approach to guidelines on therapies and for whom a more abstract interaction might provide more information. Individuals, on the other hand, are more likely to want to interact with the system on an individual level, as shown by LowID-3: "I would like to enter an individual target value for physical activity". [72] distinguish between local and global explanations of an AI system. However, to assume that end users require only local explanations would be an incorrect simplification: in fact, users express a desire to have more influence at the local level (e.g., adjusting goals for physical activity) as well as match their own calculation with the model at the global level. In any case, explanations need to be aligned and evaluated with the goals of the user.

Furthermore, our experiment shows that subjective effects may only occur after repeated interactions. Both, studies and training programs of AID-Systems, should take this effect into account. However, our results imply that, e.g., other interaction possibilities could decrease this span if necessary (c.f. [27]). AID systems should therefore ask users for their predictions in the first period AID therapy so that they can compare their own expectations with the system results with little effort. The testing of hypotheses is also a central task in order to be able to form a correct mental model about the information processing. While future studies need to investigate whether interactions with a direct goal of promoting active hypothesis testing can also increase SIPA ratings or experience traceability, it is difficult to integrate this into current AID systems. Actively inducing high or low glucose levels to compare expectations with an AID system's behavior is not recommended for medical reasons. Therefore, for XAI systems to be applicable in medical contexts such as DMT1, simulations of the algorithm need to be developed, for example, that allow this testing of hypotheses before use or as counterfactual during use. Existing approaches for the simulation of glucose level (see [100]) could be supplemented with an interface that offers explanatory variants for situations selected by the users themselves.

6.5 Limitations & Further Research

Several limitations for further research have to be considered. First, the applied method to analyze performance or prediction was not as aligned with potential tasks in real-world application as possible. That is, in AID systems users do not need to make predictions about the insulin needs calculated by the system. More importantly, they need to be able to estimate the effect of communicated information on, e.g., physical activity to cooperate effectively with the system. A more comprehensive indicator to assess the effect of traceability on the human-machine system performance could be to show a scenario and ask how changes of one or multiple attributes would affect the outcome. This would also open up different possibilities for interpretation (e.g., deviation from the correct value as in this study but also to what extent the direction of the estimate is correct as a non-metric variable). Comparable tasks exist in the area of complex problem solving [107] and could also be used in the area of human-AI interaction.

Second, in an ideal case, it would have been possible to measure the development of user experience on the course over several weeks. The time between observations, interactions and measurements in our experiment was short compared everyday application. In addition, when used in one's own therapy, one's own previous experience can be included to a greater extent. A possibility for further research could be to strive to enable longitudinal designs to allow for results based on longer reflection periods as well as personalization. In addition, participants in this experiment were shown only one condition at a time, whereas patients, for example, may compare different interfaces when deciding on an AID system. As long as the influence of learning experience is taken into account, within-subject analyses of different explanatory and interaction effects could be used in further experimental settings.

Third, the present research only examined one approach to explain to the users the way an AID system calculates insulin needs. To enable users to cope, e.g., with information overload, an interactive simulation may provide counterfactual explanations for scenarios they are interested in or want to understand. Furthermore, depending on the algorithm used to construct the AID system, the concrete depiction of rules applied to calculate insulin needs could lead to important insights into the evolution of mental models in human-AI cooperation. Ideally, further studies provide different explanations to the users in order to render it possible to compare their effectiveness for different goals (i.e., understand effects of personalization vs. understanding one own's influence on the system through communicated information).

7 CONCLUSION

Theoretically motivated and impactful research of human-centered AI is still in an early stage of development. Empirical data of potential end-users as a target group in contrast to, e.g., developers or professionals is needed. On top of that, the relationship between subjective experiences and the impact on users' capabilities to cooperate with intelligent systems is crucial for XAI applications in the future: it determines whether explanations truly empower users or, in the worst case, overburden or even deceive them. In this sense, the present work contributes to the development of human-centered XAI on three levels: By 1) refining and applying the SIPA scale, which is derived from theoretical concepts of automation, differentiated statements about the effects of explanations can be made. By 2) developing an experimental environment to examine the interaction of potential end users with AID XAI, the usefulness of explanations for everyday life can be validly assessed. And by 3) measuring performance at the same time as user experience the problematic miscalibration between perceived and actual ability to predict AI behavior can be empirically supported. Based on the empirical study, it is possible to derive design decisions that enable users of medical AI systems to collaborate and understand a system rather than overloading them with information. Future research in AID systems should therefore examine how users actively develop and test hypotheses on AID information processing to better understand under which conditions reported SIPA ratings may exhibit a better calibration with the actual task performance.

8 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).

A OVERVIEW OF ATTRIBUTES FOR AID SIMULATION

Table 11. Variables used within the AID Simulation.

Attribute	Description	Relevance
Current Tissue Glucose	The glucose level of interstitial fluid currently measured by the sensor.	It is the proxy for current blood glucose level. Needs to be in a defined range to avoid high and low blood sugar in the short term, as well as long-term problems associated with chronically high blood sugar.
Current Insulin in Body	The amount of active insulin in the body.	Lowers glucose level short term, therefore reduces the amount of insulin needed.
Current Carbohydrates in Body	The amount carbohydrates yet to be used by the body, e.g., carbohydrates in the digestive tract.	Raises glucose level (quickly or slowly depending largely on absorption rate), therefore raises the amount of insulin needed.
Current Activity	The level of physical activity of the user.	A higher activity level raises sensitivity to insulin, leads to carbohydrates being used up more quickly and thus generally lowers blood glucose, meaning it lowers the amount of insulin needed.
Tissue Glucose Target	Target amount of Glucose to be measured by the sensor as proxy for blood glucose target.	Trying to reach the blood glucose target is the primary outcome of insulin therapy for T1DM. Target value may depend on current circumstances.
Avoid Hypoglycemia	Lowers risk of low blood sugar (hypoglycemia) when activated.	Automatically reduces aggressiveness and raises glucose target, therefore reduces amount of insulin given.
Duration of Insulin Action	The time in which insulin will still be active in the body.	When insulin stays longer active or has an effect, calculations need to integrate remaining effect or effect of physical activity for remaining insulin levels .
Correction Intensity	How fast the glucose target ought to be reached. Higher aggressiveness means the glucose target ought to be reached fast.	If target glucose is below current glucose reading, high aggressiveness leads to an increased amount of insulin needed. Raises risk of hypoglycemia.
Risk of Hypoglycemia in next hour	Probability of the user experiencing hypoglycemia (low blood sugar, < 3.9 mmol/l) during the next hour.	Hypoglycemia is most likely to interfere with the user's ability to function in everyday life. A high risk of hypoglycemia therefore lets the system reduce the amount of insulin that should be given to mitigate the risk.
Blood Glucose lowering per 1 Unit Insulin	How much 1 insulin unit lowers blood glucose level. High value indicates high insulin sensitivity.	The more 1 insulin unit lowers blood glucose, the less insulin is needed.
Insulin Units per 10 grams Carbohydrates	How many insulin units need to be injected to metabolize 10 grams of carbohydrates. High value indicates low insulin sensitivity.	The more insulin units are needed to metabolize 10 grams of carbohydrates, the more insulin is needed.
Predicted Exercise	System estimate whether its expect users to exercise in the next hours.	Exercise in most cases lowers blood glucose via energy consumption and increasing insulin sensitivity. Raises glucose target automatically and thus reduces the amount of insulin given in preparation for exercise.

B DESCRIPTIVE DATA FOR ALL REPEATED MEASURES VARIABLES

Table 12. Descriptive Data for all variables measured repeatedtly at all point of measurement

		SIPA Transparency			SIPA Understandability			SIPA Predictability			FOST			NASA-TLX		
		M	*SD*	Range	*M*	*SD*	*Range*	*M*	*SD*	*Range*	*M*	*SD*	*Range*	*M*	*SD*	*Range*
Observation Block 1	LowID	5.08	1.31	4.50	4.35	1.46	5.00	3.48	1.21	4.00	4.23	1.31	4.40	4.85	1.11	4.00
	MedID	4.59	1.34	4.50	3.95	1.40	5.00	3.80	1.31	4.00	4.02	1.01	4.20	3.87	1.25	3.60
	HighID	5.10	0.82	3.00	3.90	1.13	4.00	3.90	0.96	4.00	4.33	0.86	3.40	3.57	1.25	4.60
Observation Block 2	LowID	4.40	1.40	4.50	4.04	1.47	5.00	3.35	1.25	4.50	3.90	1.32	4.40	4.82	1.36	5.00
	MedID	4.36	1.43	4.50	3.59	1.26	4.50	3.25	1.21	4.50	3.72	1.18	3.80	3.72	1.09	4.00
	HighID	5.04	0.79	2.50	3.83	0.97	4.00	3.77	1.07	4.50	4.02	1.04	4.20	3.67	1.42	5.20
Observation Block 3	LowID	4.23	1.28	5.00	3.69	1.24	5.00	3.52	1.16	5.00	3.88	1.29	4.40	4.53	1.48	6.00
	MedID	4.50	1.23	4.50	3.86	1.16	4.50	3.66	1.14	4.00	4.05	1.15	3.80	3.84	1.28	5.00
	HighID	5.02	0.87	3.00	3.94	1.35	4.50	3.67	1.50	5.00	4.13	1.30	5.00	3.84	1.43	5.20
Observation Block 4	LowID	4.23	1.36	5.00	3.77	1.32	5.00	3.35	1.13	4.50	3.75	1.41	4.60	4.67	1.43	5.80
	MedID	4.66	1.24	4.50	3.86	1.34	4.50	3.64	1.38	4.50	4.11	1.39	4.60	4.15	1.39	5.60
	HighID	5.08	0.75	3.00	4.00	1.53	5.00	3.83	1.52	5.00	4.29	1.26	4.60	4.00	1.48	5.20
Performance	LowID	4.08	1.69	5.00	3.42	1.59	5.00	3.06	1.36	4.00	3.86	1.57	4.60	3.97	1.14	4.20
	MedID	4.02	1.59	5.00	3.11	1.30	4.00	2.89	1.13	4.00	3.76	1.12	4.00	2.85	1.26	4.60
	HighID	5.02	0.83	2.50	3.98	1.36	5.00	3.77	1.31	5.00	4.42	0.90	3.60	3.41	1.30	5.40

C KNOWLEDGE QUESTIONS ON DIABETES MANAGEMENT

Table 13. Knowledge Questions on diabetes management (translated from German)

	Please indicate, whether the following statements are correct or not	True	False	I don't know
1	Even without eating, type 1 diabetics need insulin.			
2	When treating hypoglycemia, the most important goal is to get back to a level above 70 mg/dl as quickly as possible.			
3	When treating hyperglycemia, the most important goal is to get back to a level below 180 mg/dl as quickly as possible.			
4	If I am unsure of my insulin needs, I should inject too much rather than too little.			
5	Since alcohol consumption causes sugar levels to rise sharply, insulin should be administered particularly generously during a night of partying.			
6	How long insulin has an effect in the body depends, among other things, on the amount administered.			
7	"Rapid" insulin refers to insulin that takes effect immediately after injection without any delay.			
8	I can recognize increased insulin sensitivity by the fact that sugar levels drop more slowly after insulin is administered.			
9	FGM and CGM sensors measure blood glucose.			
10	The Dawn phenomenon describes how some diabetics are at high risk for hypoglycemia early in the morning (around 5 a.m.).			

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*. ACM, Montreal QC Canada, 1–18. https://doi.org/10.1145/3173574.3174156

[2] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376615

[3] Mary B. Abraham, Martin de Bock, Grant J. Smith, Julie Dart, Janice M. Fairchild, Bruce R. King, Geoffrey R. Ambler, Fergus J. Cameron, Sybil A. McAuley, Anthony C. Keech, Alicia Jenkins, Elizabeth A. Davis, David N. O'Neal, Timothy W. Jones, Australian Juvenile Diabetes Research Fund Closed-Loop Research group, Ace Choo, Jennifer Nicholas, Leah Laurenson, Alison Roberts, Keely Bebbingtom, Julie Klimek, Kristine Heels, Rebecca Gebert, Shaun Johnson, Stephanie Oats, Jordan Rafferty, Anthony Pease, Sophia Zoungas, Melissa H Lee, Barbora Paldus, Catriona M Sims, Richard J MacIssac, Glenn M Ward, Peter G Colman, Neale D Cohen, Leon Bach, Kavita Kumareswaran, Stephen N Stranks, Morton G Burt, Jane D Holmes-Walker, Roland W McCallum, Joey Kaye, Jane Speight, Christel Hendreickx, Andrzej Januszewski, Adreinne Kirby, and Sara Vogrin. 2021. Effect of a Hybrid Closed-Loop System on Glycemic and Psychosocial Outcomes in Children and Adolescents With Type 1 Diabetes: A Randomized Clinical Trial. *JAMA Pediatrics* 175, 12 (Dec. 2021), 1227. https://doi.org/10.1001/jamapediatrics.2021.3965

[4] Rebecca N. Adams, Molly L. Tanenbaum, Sarah J. Hanes, Jodie M. Ambrosino, Trang T. Ly, David M. Maahs, Diana Naranjo, Natalie Walders-Abramson, Stuart A. Weinzimer, Bruce A. Buckingham, and Korey K. Hood. 2018. Psychosocial and Human Factors During a Trial of a Hybrid Closed Loop System for Type 1 Diabetes Management. *Diabetes Technology & Therapeutics* 20, 10 (Oct. 2018), 648–653. https://doi.org/10.1089/dia.2018.0174

[5] Muhammad Aljukhadar, Sylvain Senecal, and Charles-Etienne Daoust. 2010. Information Overload and Usage of Recommendations. 612 (2010), 8.

[6] Ahlam Alotaibi, Reem Al Khalifah, and Karen McAssey. 2020. The Efficacy and Safety of Insulin Pump Therapy with Predictive Low Glucose Suspend Feature in Decreasing Hypoglycemia in Children with Type 1 Diabetes Mellitus: A Systematic Review and Meta-analysis. *Pediatric Diabetes* 21, 7 (Nov. 2020), 1256–1267. https://doi.org/10.1111/pedi.13088

[7] Yasmeen Alufaisan, Laura Ranee Marusich, Jonathan Z Bakdash, Yan Zhou, and Murat Kantarcioglu. 2020. *Does Explainable Artificial Intelligence Improve Human Decision-Making?* Preprint. PsyArXiv. https://doi.org/10.31234/osf.io/d4r9t

[8] Christiane Attig, Daniel Wessel, and Thomas Franke. 2017. Assessing Personality Differences in Human-Technology Interaction: An Overview of Key Self-Report Scales to Predict Successful Interaction. In *International Conference on Human-Computer Interaction*. Springer, 19–29.

[9] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[10] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445717

[11] Martin Baumann and Josef F Krems. 2007. Situation Awareness and Driving: A Cognitive Model. In *Modelling Driver Behaviour in Automotive Environments*. Springer, 253–265.

[12] Pierre Yves Benhamou, Erik Huneker, Sylvia Franc, Maeva Doron, and Guillaume Charpentier. 2018. Customization of Home Closed-Loop Insulin Delivery in Adult Patients with Type 1 Diabetes, Assisted with Structured Remote Monitoring: The Pilot WP7 Diabeloop Study. *Acta Diabetologica* 55, 6 (June 2018), 549–556. https://doi.org/10.1007/s00592-018-1123-1

[13] Yoav Benjamini and Henry Braun. 2002. John W. Tukey's Contributions to Multiple Comparisons. *The Annals of Statistics* 30, 6 (2002), 1576–1594.

[14] Cari Berget, Halis Kaan Akturk, Laurel H. Messer, Timothy Vigers, Laura Pyle, Janet Snell-Bergeon, Kimberly A. Driscoll, and Gregory P. Forlenza. 2021. Real-world Performance of Hybrid Closed Loop in Youth, Young Adults, Adults and Older Adults with Type 1 Diabetes: Identifying a Clinical Target for Hybrid Closed-loop Use. *Diabetes, Obesity and Metabolism* 23, 9 (Sept. 2021), 2048–2057. https://doi.org/10.1111/dom.14441

[15] Cari Berget, Laurel H. Messer, Tim Vigers, Brigitte I. Frohnert, Laura Pyle, R. Paul Wadwa, Kimberly A. Driscoll, and Gregory P. Forlenza. 2020. Six Months of Hybrid Closed Loop in the Real-world: An Evaluation of Children and Young Adults Using the 670G System. *Pediatric Diabetes* 21, 2 (March 2020), 310–318. https://doi.org/10.1111/pedi.12962

[16] Ruth Beyth-Marom and Baruch Fischhoff. 1983. Diagnosticity and Pseudodiagnosticity. *Journal of Personality and Social Psychology* 45, 6 (1983), 1185–1195. https://doi.org/10.1037/0022-3514.45.6.1185

[17] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Virtual Event USA, 401–413. https://doi.org/10.1145/3461702.3462571

[18] Alessandro Bisio, Linda Gonder-Frederick, Ryan McFadden, Daniel Cherñavvsky, Mary Voelmle, Michael Pajewski, Pearl Yu, Heather Bonner, and Sue A. Brown. 2021. The Impact of a Recently Approved Automated Insulin Delivery System on Glycemic, Sleep, and Psychosocial Outcomes in Older Adults With Type 1 Diabetes: A Pilot Study. *Journal of Diabetes Science and Technology* (Jan. 2021), 193229682098687. https://doi.org/10.1177/1932296820986879

[19] Emanuele Bosi, Pratik Choudhary, Harold W de Valk, Sandrine Lablanche, Javier Castañeda, Simona de Portu, Julien Da Silva, Roseline Ré, Linda Vorrink-de Groot, John Shin, Francine R Kaufman, Ohad Cohen, Andrea Laurenzi, Amelia Caretto, David Slatterly, Marcia Henderson-Wilson,

S. John Weisnagel, Marie-Christine Dubé, Valérie-Ève Julien, Roberto Trevisan, Giuseppe Lepore, Rosalia Bellante, Irene Hramiak, Tamara Spaic, Marsha Driscoll, Sophie Borot, Annie Clergeot, Lamia Khiat, Peter Hammond, Sutapa Ray, Laura Dinning, Giancarlo Tonolo, Alberto Manconi, Maura Serena Ledda, Wendela de Ranitz, Bianca Silvius, Anne Wojtusciszyn, Anne Farret, Titia Vriesendorp, Folkje Immeker-de Jong, Joke van der Linden, Huguette S. Brink, Marije Alkemade, Pauline Schaepelynck-Belicar, Sébastien Galie, Clémence Tréglia, Pierre-Yves Benhamou, Myriam Haddouche, Roel Hoogma, Lalantha Leelarathna, Angel Shaju, and Linda James. 2019. Efficacy and Safety of Suspend-before-Low Insulin Pump Technology in Hypoglycaemia-Prone Adults with Type 1 Diabetes (SMILE): An Open-Label Randomised Controlled Trial. *The Lancet Diabetes & Endocrinology* 7, 6 (June 2019), 462–472. https://doi.org/10.1016/S2213-8587(19)30150-0

[20] Charlotte K Boughton. 2021. Fully Closed-Loop Insulin Delivery—Are We Nearly There Yet? *The Lancet Digital Health* 3, 11 (Nov. 2021), e689–e690. https://doi.org/10.1016/S2589-7500(21)00218-1

[21] Charlotte K. Boughton, Sara Hartnell, Janet M. Allen, Julia Fuchs, and Roman Hovorka. 2022. Training and Support for Hybrid Closed-Loop Therapy. *Journal of Diabetes Science and Technology* 16, 1 (Jan. 2022), 218–223. https://doi.org/10.1177/1932296820955168

[22] Frank A. Buckless and Sue Pickard Ravenscroft. 1990. Contrast Coding: A Refinement of ANOVA in Behavioral Analysis. *The Accounting Review* 65, 4 (1990), 933–945.

[23] Ruth M. J. Byrne. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Macao, China, 6276–6282. https://doi.org/10.24963/ijcai.2019/876

[24] John T. Cacioppo and Richard E. Petty. 1982. The Need for Cognition. *Journal of Personality and Social Psychology* 42, 1 (Jan. 1982), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

[25] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300234

[26] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E Smith, and Subbarao Kambhampati. 2019. Explicability? Legibility? Predictability? Transparency? Privacy? Security? The Emerging Landscape of Interpretable Agent Behavior. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 29. 86–96.

[27] 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

[28] 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

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

[30] Jacob Cohen. 1992. Statistical Power Analysis. *Current directions in psychological science* 1, 3 (1992), 98–101.

[31] Hai Dang, Lukas Mecke, and Daniel Buschek. 2022. GANSlider: How Users Control Generative Models for Images Using Multiple Sliders with and without Feedforward Information. *arXiv:2202.00965 [cs]* (Feb. 2022). https://doi.org/10.1145/3491102.3502141 arXiv:2202.00965 [cs]

[32] Mustafa Demir, Nathan J. McNeese, and Nancy J. Cooke. 2019. The Evolution of Human-Autonomy Teams in Remotely Piloted Aircraft Systems Operations. *Frontiers in Communication* 4 (Sept. 2019), 50. https://doi.org/10.3389/fcomm.2019.00050

[33] L. Dowling, E. G. Wilmot, and P. Choudhary. 2020. Do-it-yourself Closed-loop Systems for People Living with Type 1 Diabetes. *Diabetic Medicine* 37, 12 (Dec. 2020), 1977–1980. https://doi.org/10.1111/dme.14321

[34] Jeff Druce, James Niehaus, Vanessa Moody, David Jensen, and Michael L. Littman. 2021. Brittle AI, Causal Confusion, and Bad Mental Models: Challenges and Successes in the XAI Program. *arXiv:2106.05506 [cs]* (June 2021). arXiv:2106.05506 [cs]

[35] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. 2021. The who in explainable AI: how AI background shapes perceptions of AI explanations. *arXiv preprint arXiv:2107.13509* (2021).

[36] 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

[37] 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

[38] Mica R. Endsley. 1995. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37, 1 (March 1995), 32–64. https://doi.org/10.1518/001872095779049543

[39] Mica R. Endsley and David B. Kaber. 1999. Level of Automation Effects on Performance, Situation Awareness and Workload in a Dynamic Control Task. *Ergonomics* 42, 3 (March 1999), 462–492. https://doi.org/10.1080/001401399185595

[40] C. Farrington. 2018. Psychosocial Impacts of Hybrid Closed-Loop Systems in the Management of Diabetes: A Review. *Diabetic Medicine* 35, 4 (April 2018), 436–449. https://doi.org/10.1111/dme.13567

[41] Holger Finger, Caspar Goeke, Dorena Diekamp, Kai Standvoß, and Peter König. 2017. LabVanced: A Unified JavaScript Framework for Online Studies. In *International Conference on Computational Social Science (Cologne)*.

[42] Erin D. Foster and Ariel Deardorff. 2017. Open Science Framework (OSF). *Journal of the Medical Library Association* 105, 2 (April 2017). https://doi.org/10.5195/jmla.2017.88

[43] 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

[44] David C. Funder and Daniel J. Ozer. 2019. Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science* 2, 2 (June 2019), 156–168. https://doi.org/10.1177/2515245919847202

[45] Jose Garcia-Tirado, John P. Corbett, Dimitri Boiroux, John Bagterp Jørgensen, and Marc D. Breton. 2019. Closed-Loop Control with Unannounced Exercise for Adults with Type 1 Diabetes Using the Ensemble Model Predictive Control. *Journal of Process Control* 80 (Aug. 2019), 202–210. https://doi.org/10.1016/j.jprocont.2019.05.017

[46] Ella Glikson and Anita Williams Woolley. 2020. Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals* 14, 2 (July 2020), 627–660. https://doi.org/10.5465/annals.2018.0057

[47] Jamie C. Gorman, Nancy J. Cooke, and Jennifer L. Winner. 2006. Measuring Team Situation Awareness in Decentralized Command and Control Environments. *Ergonomics* 49, 12-13 (Oct. 2006), 1312–1325. https://doi.org/10.1080/00140130600612788

[48] Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. 2019. Guidelines for Reinforcement Learning in Healthcare. *Nature Medicine* 25, 1 (Jan. 2019), 16–18. https://doi.org/10.1038/s41591-018-0310-5

[49] 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*. Vol. 52. Elsevier, 139–183.

[50] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 3 (May 2015), 407–434. https://doi.org/10.1177/0018720814547570

[51] 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). arXiv:1812.04608 [cs]

[52] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User Trust in Intelligent Systems: A Journey Over Time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM, Sonoma California USA, 164–168. https://doi.org/10.1145/2856767.2856811

[53] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.

[54] Andreas Holzinger, André Carrington, and Heimo Müller. 2019. Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. *arXiv:1912.09024 [cs]* (Dec. 2019). arXiv:1912.09024 [cs]

[55] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 624–635. https://doi.org/10.1145/3442188.3445923

[56] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (March 2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

[57] Matthew Johnson, Jeffrey M. Bradshaw, Paul J. Feltovich, Catholijn M. Jonker, M. Birna Van Riemsdijk, and Maarten Sierhuis. 2014. Coactive Design: Designing Support for Interdependence in Joint Activity. *Journal of Human-Robot Interaction* 3, 1 (March 2014), 43. https://doi.org/10.5898/JHRI.3.1.Johnson

[58] Philip Nicholas Johnson-Laird. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Number 6. Harvard University Press.

[59] Mary Joyce and Jurek Kirakowski. 2013. Development of a General Internet Attitude Scale. In *Design, User Experience, and Usability. Design Philosophy, Methods, and Tools*, Aaron Marcus (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 303–311.

[60] Barbara Kimbell, David Rankin, Nicole L. Ashcroft, Lidiya Varghese, Janet M. Allen, Charlotte K. Boughton, Fiona Campbell, Atrayee Ghatak, Tabitha Randell, Rachel E.J. Besser, Nicola Trevelyan, Roman Hovorka, Julia Lawton, and on Behalf of the CLOuD Consortium. 2020. What Training, Support, and Resourcing Do Health Professionals Need to Support People Using a Closed-Loop System? A Qualitative Interview Study with Health Professionals Involved in the Closed Loop from Onset in Type 1 Diabetes (CLOuD) Trial. *Diabetes Technology & Therapeutics* 22, 6 (June 2020), 468–475. https://doi.org/10.1089/dia.2019.0466

[61] 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

[62] Rex B Kline. 2004. Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research. (2004).

[63] Christine Knoll, Sofia Peacock, Mandy Wäldchen, Drew Cooper, Simran Kaur Aulakh, Klemens Raile, Sufyan Hussain, and Katarina Braune. 2021. Real-world Evidence on Clinical Outcomes of People with Type 1 Diabetes Using Open-source and Commercial Automated Insulin Dosing Systems: A Systematic Review. *Diabetic Medicine* (Nov. 2021). https://doi.org/10.1111/dme.14741

[64] Boris Kovatchev, Peiyao Cheng, Stacey M. Anderson, Jordan E. Pinsker, Federico Boscari, Bruce A. Buckingham, Francis J. Doyle, Korey K. Hood, Sue A. Brown, Marc D. Breton, Daniel Chernavvsky, Wendy C. Bevier, Paige K. Bradley, Daniela Bruttomesso, Simone Del Favero, Roberta Calore, Claudio Cobelli, Angelo Avogaro, Trang T. Ly, Satya Shanmugham, Eyal Dassau, Craig Kollman, John W. Lum, Roy W. Beck, and for the Control to Range Study Group. 2017. Feasibility of Long-Term Closed-Loop Control: A Multicenter 6-Month Trial of 24/7 Automated Insulin Delivery. *Diabetes Technology & Therapeutics* 19, 1 (Jan. 2017), 18–24. https://doi.org/10.1089/dia.2016.0333

[65] Joshua A. Kroll. 2021. Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 758–771. https://doi.org/10.1145/3442188.3445937

[66] 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

[67] B. Kulzer, L. Heinemann, and Timm Roos. 2021. Informationstechnologie in der Diabetesbehandlung – Erleben der Patienten. *Der Diabetologe* 17, 3 (May 2021), 265–274. https://doi.org/10.1007/s11428-021-00753-9

[68] Richard E. Ladner. 2015. Design for User Empowerment. *Interactions* 22, 2 (Feb. 2015), 24–29. https://doi.org/10.1145/2723869

[69] Sebastian Lapuschkin, Alexander Binder, Gregoire Montavon, Klaus-Robert Muller, and Wojciech Samek. 2016. The LRP Toolbox for Artificial Neural Networks. (2016), 5.

[70] David M Levy. 2008. Information Overload. *The handbook of information and computer ethics* (2008), 497.

[71] Dana Lewis. 2021. How It Started, How It Is Going: The Future of Artificial Pancreas Systems (Automated Insulin Delivery Systems). *Journal of Diabetes Science and Technology* 15, 6 (Nov. 2021), 1258–1261. https://doi.org/10.1177/19322968211027558

[72] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–15. https://doi.org/10.1145/3313831.3376590

[73] L. M. E. Lindner, W. Rathmann, and J. Rosenbauer. 2018. Inequalities in Glycaemic Control, Hypoglycaemia and Diabetic Ketoacidosis According to Socio-Economic Status and Area-Level Deprivation in Type 1 Diabetes Mellitus: A Systematic Review. *Diabetic Medicine* 35, 1 (Jan. 2018), 12–32. https://doi.org/10.1111/dme.13519

[74] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence* 2, 1 (Jan. 2020), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[75] Maria Madsen and Shirley Gregor. 2000. Measuring Human-Computer Trust. In *11th Australasian Conference on Information Systems*, Vol. 53. 6–8.

[76] Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. 2021. The Role of Explainability in Creating Trustworthy Artificial Intelligence for Health Care: A Comprehensive Survey of the Terminology, Design Choices, and Evaluation Strategies. *Journal of Biomedical Informatics* 113 (2021), 103655.

[77] John M. McGuirl and Nadine B. Sarter. 2006. Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 48, 4 (Dec. 2006), 656–665. https://doi.org/10.1518/001872006779166334

[78] Stephanie M. Merritt, Deborah Lee, Jennifer L. Unnerstall, and Kelli Huber. 2015. Are Well-Calibrated Users Effective Users? Associations Between Calibration of Trust and Performance on an Automation-Aided Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 1 (Feb. 2015), 34–47. https://doi.org/10.1177/0018720814561675

[79] Laurel H. Messer, Cari Berget, Tim Vigers, Laura Pyle, Cristy Geno, R. Paul Wadwa, Kimberly A. Driscoll, and Gregory P. Forlenza. 2020. Real World Hybrid Closed-loop Discontinuation: Predictors and Perceptions of Youth Discontinuing the 670G System in the First 6 Months. *Pediatric Diabetes* 21, 2 (March 2020), 319–327. https://doi.org/10.1111/pedi.12971

[80] Laurel H. Messer, Gregory P. Forlenza, Jennifer L. Sherr, R. Paul Wadwa, Bruce A. Buckingham, Stuart A. Weinzimer, David M. Maahs, and Robert H. Slover. 2018. Optimizing Hybrid Closed-Loop Therapy in Adolescents and Emerging Adults Using the MiniMed 670G System. *Diabetes Care* 41, 4 (April 2018), 789–796. https://doi.org/10.2337/dc17-1682

[81] Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv:1706.07269 [cs]* (Aug. 2018). arXiv:1706.07269 [cs]

[82] Majid Mobasseri, Masoud Shirmohammadi, Tarlan Amiri, Nafiseh Vahed, and Hossein Hosseini Fard. 2020. Prevalence and Incidence of Type 1 Diabetes in the World: A Systematic Review and Meta-Analysis. (2020), 18.

[83] Mobeen Nazar, Muhammad Mansoor Alam, Eiad Yafi, and Mazliham Mohd Su'ud. 2021. A Systematic Review of Human–Computer Interaction and Explainable Artificial Intelligence in Healthcare With Artificial Intelligence Techniques. *IEEE Access* 9 (2021), 153316–153348. https://doi.org/10.1109/ACCESS.2021.3127881

[84] Minh Nguyen, Ivana Jankovic, Laurynas Kalesinskas, Michael Baiocchi, and Jonathan H Chen. 2021. Machine Learning for Initial Insulin Estimation in Hospitalized Patients. *Journal of the American Medical Informatics Association* 28, 10 (Sept. 2021), 2212–2219. https://doi.org/10.1093/jamia/ocab099

[85] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces*. 340–350.

[86] Emil Øversveen. 2020. Stratified Users and Technologies of Empowerment: Theorising Social Inequalities in the Use and Perception of Diabetes Self-management Technologies. *Sociology of Health & Illness* 42, 4 (May 2020), 862–876. https://doi.org/10.1111/1467-9566.13066

[87] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39, 2 (June 1997), 230–253. https://doi.org/10.1518/001872097778543886

[88] 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 (May 2000), 286–297. https://doi.org/10.1109/3468.844354

[89] John Patrick and Philip L. Morgan. 2010. Approaches to Understanding, Analysing and Developing Situation Awareness. *Theoretical Issues in Ergonomics Science* 11, 1-2 (Jan. 2010), 41–57. https://doi.org/10.1080/14639220903009946

[90] Margus Pedaste, Mario Mäeots, Leo A Siiman, Ton De Jong, Siswa AN Van Riesen, Ellen T Kamp, Constantinos C Manoli, Zacharias C Zacharia, and Eleftheria Tsourlidaki. 2015. Phases of Inquiry-Based Learning: Definitions and the Inquiry Cycle. *Educational research review* 14 (2015), 47–61.

[91] Peter Pesl, Pau Herrero, Monika Reddy, Maria Xenou, Nick Oliver, Desmond Johnston, Christofer Toumazou, and Pantelis Georgiou. 2016. An Advanced Bolus Calculator for Type 1 Diabetes: System Architecture and Usability Results. *IEEE Journal of Biomedical and Health Informatics* 20, 1 (Jan. 2016), 11–17. https://doi.org/10.1109/JBHI.2015.2464088

[92] Barbara Piccini, Emilio Casalini, Chiara Macucci, and Sonia Toni. 2022. Type 1 Diabetes Technology Management Traps in a Pediatric Patient: Not All That Glitters Is Gold. *Acta Diabetologica* 59, 1 (Jan. 2022), 137–141. https://doi.org/10.1007/s00592-021-01781-z

[93] 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

[94] David Rankin, Barbara Kimbell, Janet M. Allen, Rachel E. J. Besser, Charlotte K. Boughton, Fiona Campbell, Daniela Elleri, Julia Fuchs, Atrayee Ghatak, Tabitha Randell, Ajay Thankamony, Nicola Trevelyan, Malgorzata E. Wilinska, Roman Hovorka, and Julia Lawton. 2021. Adolescents' Experiences of Using a Smartphone Application Hosting a Closed-loop Algorithm to Manage Type 1 Diabetes in Everyday Life: Qualitative Study. *Journal of Diabetes Science and Technology* 15, 5 (Sept. 2021), 1042–1051. https://doi.org/10.1177/1932296821994201

[95] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat]* (Feb. 2016). arXiv:1602.04938 [cs, stat]

[96] Lysanne Rivard, Pascale Lehoux, and Hassane Alami. 2021. "It's Not Just Hacking for the Sake of It": A Qualitative Study of Health Innovators' Views on Patient-Driven Open Innovations, Quality and Safety. *BMJ Quality & Safety* 30, 9 (Sept. 2021), 731–738. https://doi.org/10.1136/bmjqs-2020-011254

[97] Ralph L. Rosnow, Robert Rosenthal, and Donald B. Rubin. 2000. Contrasts and Correlations in Effect-Size Estimation. *Psychological Science* 11, 6 (Nov. 2000), 446–453. https://doi.org/10.1111/1467-9280.00287

[98] 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.

[99] Swati Sachan, Jian-Bo Yang, Dong-Ling Xu, David Eraso Benavides, and Yang Li. 2020. An Explainable AI Decision-Support-System to Automate Loan Underwriting. *Expert Systems with Applications* 144 (April 2020), 113100. https://doi.org/10.1016/j.eswa.2019.113100

[100] Jana Schmitzer, Carolin Strobel, Ronald Blechschmidt, Adrian Tappe, and Heiko Peuscher. 2022. Efficient Closed Loop Simulation of Do-It-Yourself Artificial Pancreas Systems. *Journal of Diabetes Science and Technology* 16, 1 (2022), 61–69.

[101] 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. In *Proceedings of the ACM CHI Workshop on Operationalizing Human-Centered Perspectives in Explainable AI (HCXAI 2021)*, Upol Ehsan, Q. Vera Liao, Martina Mara, Mark Riedl, Andreas Riener, Marc Streit, Sandra Wachter, and Philipp Wintersberger (Eds.). The Internet.

[102] Peter Sedlmeier and Frank Renkewitz. 2018. *Forschungsmethoden und Statistik in der Psychologie* (nachdr. ed.). Pearson Studium, München.

[103] Mark R Segal. 2004. Machine Learning Benchmarks and Random Forest Regression. (2004).

[104] Clare Shaban. 2015. Psychological Themes That Influence Self-Management of Type 1 Diabetes. *World Journal of Diabetes* 6, 4 (2015), 621. https://doi.org/10.4239/wjd.v6.i4.621

[105] Joseph P. Shivers, Linda Mackowiak, Henry Anhalt, and Howard Zisser. 2013. "Turn It Off!": Diabetes Device Alarm Fatigue Considerations for the Present and the Future. *Journal of Diabetes Science and Technology* 7, 3 (May 2013), 789–794. https://doi.org/10.1177/193229681300700324

[106] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (April 2020), 495–504. https://doi.org/10.1080/10447318.2020.1741118

[107] Valerie J. Shute, Lubin Wang, Samuel Greiff, Weinan Zhao, and Gregory Moore. 2016. Measuring Problem Solving Skills via Stealth Assessment in an Engaging Video Game. *Computers in Human Behavior* 63 (Oct. 2016), 106–117. https://doi.org/10.1016/j.chb.2016.05.047

[108] Madison B. Smith, Anastasia Albanese-O'Neill, Tamara G.R. Macieira, Yingwei Yao, Joseph M. Abbatematteo, Debra Lyon, Diana J. Wilkie, Michael J. Haller, and Gail M. Keenan. 2019. Human Factors Associated with Continuous Glucose Monitor Use in Patients with Diabetes: A Systematic Review. *Diabetes Technology & Therapeutics* 21, 10 (Oct. 2019), 589–601. https://doi.org/10.1089/dia.2019.0136

[109] Aaron Springer and Steve Whittaker. 2018. "I Had a Solid Theory before but It's Falling Apart": Polarizing Effects of Algorithmic Transparency. *arXiv:1811.02163 [cs]* (Nov. 2018). arXiv:1811.02163 [cs]

[110] Aaron Springer and Steve Whittaker. 2020. Progressive Disclosure: When, Why, and How Do Users Want Algorithmic Transparency Information? *ACM Transactions on Interactive Intelligent Systems* 10, 4 (Dec. 2020), 1–32. https://doi.org/10.1145/3374218

[111] Jackie Sturt, Kathryn Dennick, Mette Due-Christensen, and Kate McCarthy. 2015. The Detection and Management of Diabetes Distress in People With Type 1 Diabetes. *Current Diabetes Reports* 15, 11 (Nov. 2015), 101. https://doi.org/10.1007/s11892-015-0660-z

[112] Sakinah Suttiratana, Jessie J. Wong, Monica S. Lanning, Adrienne Dunlap, Sarah Hanes, Korey K. Hood, Rayhan Lal, and Diana Naranjo. 2021. 518-P: User Experiences with Loop, an Open-Source Automated Insulin Delivery (AID) System. *Diabetes* 70, Supplement_1 (June 2021), 518–P. https://doi.org/10.2337/db21-518-P

[113] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, Textual or Hybrid: The Effect of User Expertise on Different Explanations. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 109–119. https://doi.org/10.1145/3397481.3450662

[114] R. M. Taylor. 1989. Situational Awareness Rating Technique (SART): The Development of a Tool for Aircrew Systems Design. In *Proceedings of the AGARD AMP Symposium on Situational Awareness in Aerospace Operations, CP478*. Seuilly-sur Seine: NATO AGARD.

[115] Sara Trevitt, Sue Simpson, and Annette Wood. 2016. Artificial Pancreas Device Systems for the Closed-Loop Control of Type 1 Diabetes: What Systems Are in Development? *Journal of Diabetes Science and Technology* 10, 3 (May 2016), 714–723. https://doi.org/10.1177/1932296815617968

[116] Daniel Trommler, Christiane Attig, and Thomas Franke. 2018. Trust in Activity Tracker Measurement and Its Link to User Acceptance. *Mensch und Computer 2018-Tagungsband* (2018).

[117] Silvia Tulli, Filipa Correia, Samuel Mascarenhas, Samuel Gomes, Francisco S. Melo, and Ana Paiva. 2019. Effects of Agents' Transparency on Teamwork. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, Davide Calvaresi, Amro Najjar, Michael Schumacher, and Kary Främling (Eds.). Springer International Publishing, Cham, 22–37.

[118] 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*. ACM, Glasgow Scotland Uk, 1–15. https://doi.org/10.1145/3290605.3300831

[119] Kathryn W. Weaver and Irl B. Hirsch. 2018. The Hybrid Closed-Loop System: Evolution and Practical Applications. *Diabetes Technology & Therapeutics* 20, S2 (June 2018), S2–16–S2–23. https://doi.org/10.1089/dia.2018.0091

[120] Bert Weijters and Hans Baumgartner. 2012. Misresponse to Reversed and Negated Items in Surveys: A Review. *Journal of Marketing Research* 49, 5 (Oct. 2012), 737–747. https://doi.org/10.1509/jmr.11.0368

[121] Christopher D. Wickens and C. Melody Carswell. 2006. Information Processing. In *Handbook of Human Factors and Ergonomics*, Gavriel Salvendy (Ed.). John Wiley & Sons, Inc., Hoboken, NJ, USA, 111–149. https://doi.org/10.1002/0470048204.ch5

[122] Stefan Wiens and Mats E. Nilsson. 2017. Performing Contrast Analysis in Factorial Designs: From NHST to Confidence Intervals and Beyond. *Educational and Psychological Measurement* 77, 4 (Aug. 2017), 690–715. https://doi.org/10.1177/0013164416668950

[123] Tao Zhang, David Kaber, and Maryam Zahabi. 2022. Using Situation Awareness Measures to Characterize Mental Models in an Inductive Reasoning Task. *Theoretical Issues in Ergonomics Science* 23, 1 (Jan. 2022), 80–103. https://doi.org/10.1080/1463922X.2021.1885083

[124] 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

[125] 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