What Clinicians Need: Designing, Developing and Evaluating an AI-Based Decision Support System for Autism Assessment

ANONYMOUS AUTHOR(S)

AI methods promise to support autism spectrum condition (ASC) diagnostics in adults, a complex and time-consuming process, that is characterized by a shortage of specialized clinicians. To date, clinicians' needs and their interaction with such AI-based support remain underexplored. Our work aims to develop and evaluate an AI-based clinical decision support system (CDSS) for ASC assessment, and to investigate how it impacts clinicians' decision-making. By interviewing clinicians of varying experience levels, we identified five challenges and derived design strategies. Based on that, we developed SIT-CARE, a CDSS, which provides AI-based recommendations and data visualizations of clinically relevant non-verbal behavior. Through an evaluation study with newly recruited clinicians, we found that SIT-CARE led to different decision paths in regard to the ASC assessment, which are reflected in clinicians' mental models. Overall, SIT-CARE demonstrated potential in supporting the complex process of ASC assessments.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; • **Applied computing** → **Health informatics**; • **Computing methodologies** → *Artificial intelligence*.

Additional Key Words and Phrases: Human-AI collaboration, Medical decision making, Autism Spectrum Conditions, Mental Models

ACM Reference Format:

Anonymous Author(s). 2018. What Clinicians Need: Designing, Developing and Evaluating an AI-Based Decision Support System for Autism Assessment. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 28 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Artificial Intelligence (AI) and machine learning (ML) have shown significant potential in improving diagnostic accuracy across various medical domains (e.g., [17, 41]). In this work, we focus specifically on AI-assistance for the diagnosis process of Autism Spectrum Conditions (ASC) in adults.

Autism Spectrum Condition (ASC) is recognized as a neurodevelopmental condition characterized by differences in social interaction and communication, as well as patterns of restricted or repetitive behavior [6, 56]. According to the Centers for Disease Control and Prevention (CDC), approximately 3% of the global population has ASC [7, 83]. Further, research indicates a high prevalence of undiagnosed ASC among adults with average or higher intelligence [42, 72]. This may be due to the complexity of the ASC diagnosis process in adults. ASC assessments currently used are time-consuming, require extensive training, only have moderate accuracy in adults on their own [22], and often rely on subjective self-reports of patients [83]. Therefore, clinicians require many years of training and experience to accurately diagnose ASC. Furthermore, due to a shortage of specialized clinicians [43, 73], the high demand for ASC diagnostics can not be met [23]. To address these challenges, research has begun to explore the potential of AI to support ASC assessments (see [33, 36]). Especially, AI trained on more objective, behavioral data seems to be promising [38]. To do so, such behavioral data first has to be collected under reproducible conditions. For this the Simulation Interaction

Manuscript submitted to ACM

Task (SIT) [24] can be used. The SIT is a standardized task to support ASC assessment in adults, in which patients interact with a video-based interviewer about their likes and dislikes, enabling the collection of video data during the interaction. Prior work shows, that information of patients' verbal and non-verbal behavior can be extracted from the SIT video data and be used for AI model training [24, 61].

To the best of our knowledge, AI models using reproducible data have not yet been integrated into a clinical decision support system (CDSS). **To bridge this gap, we designed, developed and evaluated an AI-based CDSS that focuses on ASC in adults and uses objective data non-verbal behavioral data collected under reproducible conditions with the SIT.** However, recent research suggests that assisting decision-making with AI does rarely lead to improved human-AI team performance [70], underlining the need to develop **human-centered AI (HCAI)** by considering the task at hand and involving users in the process [19, 70]. Thus, for our work, clinicians are involved in the design process to consider the clinicians' needs, answering our first research question: "*What are challenges clinicians face in their decision-making process of diagnosing ASC in adults, and how can they be supported with AI in their process?*". For a human-centered AI, the task at hand, i.e., medical decision-making [64], needs to be considered and clinicians are needed to evaluate the system [19, 29]. To evaluate not only if but how an AI-based CDSS influences the clinicians' decision-making, we need to investigate their mental models during this continuous process, i.e., simplified and incomplete understanding of how an AI system works and behaves [10, 37, 54]. Thus, our second research question is: "*How does our AI-based CDSS for ASC assessment support for non-verbal behavior influence clinicians decision-making process, and their mental models?*"

To design, develop and evaluate our AI-based CDSS for ASC assessments, SIT-CARE, we first interviewed clinicians of varying experience levels to get an insight into their needs and diagnostic workflow. The interviews revealed that clinicians want AI-based support and led to five challenges with respective design strategies to apply in an implementation, thus answering RQ1. By considering these five design strategies, we developed the AI-based SIT-CARE. The SIT-CARE user interface consists of the DATA-BASED ASSESSMENT, which provides an insight into selected non-verbal behavior extracted from the SIT recording in form of visualizations and additional information, to support earlier decision-making stages [64, 82]. Further, the MODEL-BASED ASSESSMENT provides an AI recommendation accompanied by information about the model and prediction to support intermediate and later decision-making stages [64, 82]. Finally, we conducted a user study to explore how the SIT-CARE influenced clinicians' decision-making. Clinicians assessed two patient cases. They assessed each case three times, i.e., first based on the SIT video, then again after seeing the DATA-BASED ASSESSMENT and lastly, after exploring the MODEL-BASED ASSESSMENT. This allowed us to identify different decision paths, which are reflected in the clinicians' mental models and characterized by changing decisions after receiving information from the SIT-CARE. Overall, all clinicians were in favor of the SIT-CARE, and found it to be a valuable ASC screening support[1] and learning opportunity.

This paper demonstrates the following contributions:

- Through a formative study with clinicians, we explored the clinicians' workflow and challenges in the diagnosis process of ASC, and identified how an AI-based CDSS has to be designed to support them in their process.
- We developed a new AI-based CDSS, SIT-CARE, which consists of the DATA-BASED ASSESSMENT to inform clinicians about a patient's non-verbal behavior, and the MODEL-BASED ASSESSMENT to recommend whether further ASC assessments may be needed.

[1] A screening is a brief initial assessment before an in-detail diagnosis is considered.

- In an user evaluation study, we investigated how clinicians interact with the SIT-CARE, and identified different decision paths, which are reflected in clinicians' mental models.

2 Related Work

2.1 Autism Spectrum Condition Diagnostics in Adults

Currently, ASC diagnostics rely on behavioral and observational assessments due to the lack of conclusive biomarkers [83]. For a brief ASC screening, questionnaires such as the AQ-10 questionnaire [2] and the Ritvo Autism and Asperger Diagnostic Scale-Revised (RAADS-R [59]) are used to flag adults for referral to a specialist. However, they rely on self-reporting and their utility is not undisputed [5, 22, 32]. Thus, the current standard for diagnosing adults with ASC is a semi-structured assessment, i.e., the Autism Diagnostic Observation Schedule (ADOS [45]) or the Autism Diagnostic Interview–Revised (ADI-R [46]), which is conducted by a specialist and does not rely on self-reporting. Nevertheless, these instruments are time-consuming, require extensive training, and only have moderate accuracy in adults on their own [22]. Conner et al. [22] found that even the gold standard ADOS had only 65% sensitivity in an outpatient sample, indicating the difficulty of diagnosing ASC in adults. In addition, adults with higher intelligence often develop coping and masking strategies[2] that can complicate clinicians' assessment [26, 42, 72]. These challenges of traditional methods led to a growing interest in using AI to support and improve the ASC diagnostic and screening process.

2.2 AI-Supported Diagnostics for Autism Spectrum Conditions

AI has been applied to adult ASC screening questionnaires, such as the AQ-10 and RAADS-R (e.g., [67]). However, training AI with subjective self-reported experiences can be problematic, as these reports can be unreliable or incomplete, particularly in adults who use compensatory strategies. Other AI-based approaches focus on behavioral information similar to current diagnostic assessments like the ADOS. For example, Kumar and Jaiswal [38] have used AI techniques to identify subsets of behavioral features from the ADOS that achieved high sensitivity and specificity, comparable to the AI model considering all ADOS features. Another promising approach is the automated analysis of standard video recordings [66]. Overall, AI models have shown strong potential in detecting subtle and complex behavioral cues, such as facial expressions [14, 38], gaze dynamics [35], and gesture patterns add source— that may signal autistic traits and could be missed by standard clinical observation.

Currently, few AI approaches for ASC in adults have progressed from experimental validation to practical, scalable diagnostic tools [12, 20]. Thus, there is a need for AI-based CDSS that can be integrated into routine procedures for ASC diagnostics, enabling structured, multimodal observation of social-communicative behavior in adults. To train such an AI model, social-communicative behavior must be collected under reproducible conditions. For this, the Simulated Interaction Task (SIT) [24, 61], a computer-based social interaction paradigm designed to elicit verbal and non-verbal behavior under standardized conditions in adults, can be used. For example, Saakyan et al. [61] collected a dataset of 164 participants (83 with ASC), showcasing the scalability and accessibility of the SIT. Using this dataset consisting of video and audio data, the authors trained a multimodal AI model by extracting facial expressions, head movements, and vocal features, which distinguished adults with and without an ASC diagnosis (accuracy > 70%).

To date, such a reasonably accurate AI model trained on objective and reproducible SIT data has not yet been integrated into a human-centered application for clinicians to support ASC assessments.

[2]Individuals with ASC may employ camouflaging or masking strategies and behavior to conceal their autistic characteristics, with the intention of appearing "non-autistic" in social situations. [49]

2.3 Designing and Developing Clinician-centered AI

Our goal of providing clinicians with an AI-based CDSS is to support their decision making with respect to ASC assessments. However, a recent meta-analysis found that human-AI teams, i.e., humans assisted by AI, do not outperform their own or the AI's performance in decision-making tasks [70]. Highly researched approaches that "*produce details or reasons to make [AI's] functioning clear or easy to understand*", i.e., Explainable AI (XAI) [11], did not lead to improved team performance. This could be due to the technology-focused perspective "*aiming to enhance algorithmic effects rather than meet user needs*" [83]. Hence, the field of Human-Computer Interaction (HCI) has begun to shift to human-centered design approaches in order to develop human-centered artificial intelligence (HCAI) [19]. HCAI encompasses involving humans throughout the process, from design to evaluation. In addition, aspects such as the type of task should be considered more [29, 70].

Especially in the medical field, Capel and Brereton [19] highlight, human-centered approaches are often used by involving clinical specialists in the development process and considering the context of the interaction, such as the medical decision-making workflow (e.g., [60, 74, 75, 79, 80, 82]). For example, Zhang et al. [82] conducted a formative study involving clinicians to improve an existing AI-powered sepsis prediction module, followed by an evaluation study. By considering the clinicians' decision-making process, the authors found that they need to support earlier and intermediate medical decision-making stages (e.g., hypotheses generation and data gathering [64]) instead of only focusing on later stages (e.g., final decision [64]). Similarly, Wolf et al. [79] focused on earlier and intermediate medical decision-making stages, and found that their AI-based CDSS triggered analytical thinking.

Thus, to design and develop a **human-centered AI-based CDSS**, clinicians need to be involved throughout the process [19] and their needs, the type of task [29, 70], in this case medical decision-making steps [64], have to be considered.

2.4 Understanding Users' Mental Models during AI-assisted Clinical Decision-Making

To understand how clinicians use such a tool and improve human-AI collaboration, we need to understand the clinicians' reasoning during their interaction with the AI system, i.e., their *mental model*. A mental model is a simplified, incomplete understanding of how the AI behaves and works that a user develops as they interact with the AI system [10, 37, 54]. This understanding influences how users interact with, comprehend, and predict the AI system's actions [30, 34, 55]. Studies indicate that a user's mental model of an AI system influences their ability to interact with it effectively, which in turn may affect team performance [10, 15], and in our case diagnostic accuracy. Various methods can be used to explore a user's reasoning and mental model of an AI system, ranging from questionnaires (e.g., [21, 37, 40]), task performance and behavior (e.g., [10, 53]) to prediction tasks [30]. Qualitative methods like think-aloud protocols and interviews can investigate a user's mental model in more depth, revealing why they made a decision (e.g., [15, 28]). Furthermore, research investigated users' reliance on AI by considering different decision paths that humans could follow [51, 62]. For example, a user may be asked to make decisions at different *decision points*, such as at the outset, after receiving AI advice and after receiving explanations. Changes in decisions indicate shifts in the user's mental model.

To support ASC diagnostics it is relevant to align the AI-CDSS appropriately with the clinicians' workflow, thus, their medical decision-making process has to be investigated in detail [76, 81]. To do so approaches to explore a user's mental model can be combined. By conducting an interview with a compartmentalized interaction with several decision points [51], the clinicians can make a decision and predict the AI's recommendation at each point [10, 30]. In addition,

qualitative methods (e.g., [8, 80, 82]) allow us to ask questions at each decision point questions can be asked to gain an in-depth insight of a user's current mental model.

By studying clinicians' mental models during decision-making more continuously, we can gain a richer understanding of how our AI-based CDSS influences their thought process and make informed decisions about how to refine and improve the tool.

3 Formative Study: Current challenges and opinions on AI-assisted ASC screening support

To better understand the challenges clinicians encounter during the decision-making process of diagnosing ASC in adults, we conducted open-ended semi-structured interviews [44]. Through these interviews, we aimed at exploring clinicians' needs, their diagnostic workflow, and gathering feedback on how non-verbal behavior derived from the SIT video data could be effectively communicated.

3.1 Method

3.1.1 Participants, Data Collection and Analysis. We recruited seven clinicians, i.e., practicing psychotherapists, via purposive sampling [4], see Table 1.To understand the clinicians' needs, we included participants with varying levels of expertise ranged from psychotherapists in training to practicing psychotherapists who are specialized in ASC. Three of the authors were present for each interview, which was conducted remotely via an institutional Webex platform. Each session lasted an average of 49 minutes. This study was approved by the IRB of the second author's institution. All participants provided informed consent before participating A total of six hours of audio material was recorded, transcribed, and analyzed using an inductive coding approach [48]. The first author read the transcripts multiple times to ensure familiarity, and derived the initial codes by segmenting the text into meaningful units. These codes were iteratively grouped into higher-level categories and refined into overarching themes. The findings were collaboratively structured and refined through iterative feedback and team discussions.

3.1.2 Study Procedure. During the interview, the participants were asked questions about their demographic background, their experience with ASC, and their diagnostic workflows. Next, we introduced the SIT and explained that video data of a person during the SIT can be recorded and used to analyze non-verbal behavior. We then asked participants to identify non-verbal behavior they considered most relevant for ASC diagnosis. After capturing their initial thoughts, a list of non-verbal behavior was shown, and the clinicians were asked to assess whether being provided information about which of these non-verbal behavior would be helpful. We presented four schematic visualizations of non-verbal behavior, including gaze, mimicry, voice, and head behavior, to gather feedback on their usefulness and integration. By showing examples with different plot types, i.e., line, box and bi-dimensional histogram plots, we were able to get feedback on what plot type is the most useful, liked and on how such information could be integrated into an implementation. Finally, we asked about what challenges and benefits of integrating SIT and its data visualizations into their diagnostic workflows they expect. An outline of the interview protocol is provided in Appendix A.

3.2 Findings

Our interviews revealed five overarching themes, each encompassing identified challenges and corresponding design strategies for an AI-based ASC screening support system.

3.2.1 Supporting Clinicians to Meet Diagnostic Demand. All clinicians underlined that diagnosing ASC requires years of training and experience compared to other diagnoses, as knowledge about specific alterations of non-verbal behavior

Table 1. Demographics of Participants in our Formative Study.

P#	Gender	Professional Stage	Experience
P1	Female	Psychotherapist in Training	0.5 years
P2	Female	Practicing psychotherapist	5 years, specialized in ASC
P3	Female	Practicing psychotherapist	>15 years, specialized in ASC
P4	Female	Psychotherapist in Training	3 years
P5	Female	Practicing psychotherapist	25 years, specialized in ASC
P6	Female	Practicing psychotherapist	5 years, specialized in ASC
P7	Female	Practicing psychotherapist	>25 years, specialized in ASC

in affected individuals is necessary (P5, P2, P7). Furthermore, P5 described "*We open up for new registrations every three months, [...] we had 1,400 registrations for an estimated 30 people that we can actually take on.*" Consequently, due to the high demand for autism diagnoses, there has been a need to support clinicians in their assessments and enable more clinicians to perform initial screenings (P6, P2).

To address this, P5 recommended "*to outsource certain parts of the diagnostic process*" and noted that "*non-verbal symptoms is something that clinicians who are not familiar with the field find particularly difficult.*" P2 commented that "*[AI-assisted support] could be really cool in a screening context, because then people [...] can be very well-supported in this case, and maybe even identify people who can continue with the diagnostic process.*". All participants expressed interest in AI-assisted support, but noted the need for additional guidance, such as definitions for clinical and technical terms. (P1).

In summary, due to the high demand for autism diagnosis clinicians need additional support such as AI-based assistance **(Design Strategy 1)** to improve both the screening and diagnosis process and empower less experienced clinicians.

3.2.2 Providing understandable information to foster objectivity in ASC assessments. Participants raised the concern that the assessments currently used to diagnose ASC heavily rely on patients' self-reports and clinicians' interpretations, as no "*objective*" measures, such as biomarkers, are currently established (P1, P3, P6). P7 explained, "*[The ADOS] is not very informative [especially for high functional ASC], and in the end it is just a subjective assessment by us [the clinicians]*". P2 reflected: "*Do I really recognize well, even though I've been doing this for a bit longer now, whether this person is actually very, very good at masking or if they're just good at [social interactions]?*" and explained that gender bias may influence the clinician's subjective assessment. Additionally challenging is that most patients already suspect having ASC and may disagree with a clinician's assessment (P3). In such cases, clinicians often feel pressured to provide more objective evidence (P7).

Participants emphasized that "*[to make the decision] less dependent on my subjective impression, but rather to make it somehow more objective, [...] would be fantastic*" (P2) and a "*a great additional proof.*" (P7). P4 mentioned that they consider to carefully select some visualizations representing non-verbal behavior to discuss them with patients. Summarized, all participants expressed interest in visualizations of non-verbal behavior, as long as they are understandable. By collecting feedback on the example visualizations, we were able to extract how the participants want objective data to be presented. While box plots were described as not helpful or understandable (P2, P3, P6), line plots showing information

over time were fancied (P1, P2, P4). The bi-dimensional histogram plots for gaze information was described as intuitive (P2, P6, P7), but it was recommended to illustrate the display and actress in the plot (P5).

Summarized, participants want objective information presented in an understandable way **(Design Strategy 2)**.

3.2.3 Balancing the amount and detail of information. During the interviews, we asked clinicians which non-verbal behavioral data would be of interest and presented visualization drafts. While all clinicians expressed high interest in most of the data, it was also noted that providing too much information could lead less experienced clinicians to consider irrelevant information, as P5 states: *"it's always, a bit less is more"*. Also, the worry that providing such analysis of non-verbal behavior may be to complicated was mentioned (P6).

The participants recommended to focus on non-verbal behavior that *"have been found to be the most informative in research and clinicians also observe themselves"* or *"that are in the* ICD[3]" (P5). Additional details could be presented in tables or numerical formats (P5, P7). P6 emphasized the importance of detail on demand, stating *"so that I can get important information at a glance, but more information if I want it."* Similarly, P4 suggested *'I would show all the visualizations, but not all at once [...]. you can unfold them individually, so that you're not overwhelmed by so many presentations"* while also recommending the ability to *"interactively switching between the reference groups"*.

In summary, non-verbal behavior should be selected based on clinical relevance, with adjustable information levels (**Design Strategy 3**).

3.2.4 Considering the existing clinical workflow. In the interviews, it became apparent that the diagnostic process is "*extremely time-consuming*" (P2, P6). Similarly, P1 explicitly noted that the current ASC diagnosis already consists of many different assessments and that integrating anything new into their workflow would be difficult. For example, the participants described that besides many other assessments they also judge a person's non-verbal behavior with the ADOS (P3, P6, P7). Sometimes even case conferences, supervisions or feedback from colleagues may be needed, especially for edge cases (P2, P4, P6).

Participants stated that many of the discussed non-verbal behavior are similar to aspects they are familiar with from the ADOS, and that data-based information would be a good addition (P2, P6). P4 and P6 highlighted the advantage of having standardized video recordings, that could be discussed in case conferences. Furthermore, multiple participants underlined their expertise is built on a lot of training and experience from numerous cases (P2), and that providing visualizations of example cases as an orientation would be helpful (P5, P7). In addition, participants want to be able to compare current patient data with reference values, such as from control groups without ASC (P2, P5), differentiated by gender or other diagnoses (P2, P7). Similarly, P5 described that *"[they] would like to see the reference values for people without autism in the image [...] with standard deviations or something like that*", which is similar to their current practice of comparing a patient's questionnaire score with references.

Summarized, for a seamless integration into existing diagnostic workflows we aim at building on concepts that are familiar to the clinicians **(Design Strategy 4)**.

3.2.5 Setting Expectations. Participants critically reflected on how such a tool should be introduced and how uncertainty should be communicated. Participants explained that *"this little building block [data-based support] is intended to help with non-verbal communication, this small part of the diagnosis, and to do so with a certain level of accuracy"* (P5). It should be avoided to diagnose *"based on a single symptom"* (P7) P5 further stated that *"The responsibility also lies with the developers"* to ensure that a tool is used as intended.

[3]ICD is the International Classification of Diseases.

Table 2. Challenges, Design Strategies and Implementation based on the formative study

Challenge	Design Strategy	Summary of clinicians' needs
High demand for autism diagnosis	Additional support in their diagnosis process	– Support judging non-verbal behavior – Outsourcing parts of diagnostic process – AI-assisted support in ASC screening wanted
Current assessments are subjective impressions of clinicians and influenced by the patient's own impression	Providing understandable information to foster objectivity in ASC assessments	– Present understandable visualizations non-verbal behavior – Intuitive plots (e.g., line plot, bi-dimensional histogram plot) are preferred over box plots – Gaze behavior towards the display can be illustrated directly on an illustration with a display and the actress
Too much information may be overwhelming	Balancing the amount and detail of information	– Only selected non-verbal behavior should be visualized – Additional information can be provided in text or tables – Allowing user to control the wanted amount of information with an interactive interface
Already time-consuming diagnosis process	Considering the existing clinical workflow	– Connect the new data-based information with known concepts (e.g., ADOS) – Provide visualizations depicting non-verbal behavior from example cases as a comparison – Provide and visualize reference values
Misuse of provided recommendations and information	Setting Expectations	– Communicate intended use and limitations clearly – Inform about the system's accuracy

P7 recommended to "*exactly explain [to the clinicians] what this little building block is [made] for*", and that it does not replace the diagnosis process (P5). Furthermore, participants expressed the need to know the system's accuracy, for example, the probability of the system recognizing the non-verbal behavior as typical for ASC (P5).

Summarized, users need to be informed about the intended use and system limitations (**Design Strategy 5**).

3.3 Summary of Findings

Answering our first research question, our formative study revealed challenges clinicians face in their current ASC diagnoses, as well as design strategies that could address these challenges, see Table 2. Currently, the demand for ASC diagnosis is rising, while there is a lack of specialized clinicians who are able to conduct this subjective and very time-consuming diagnosis process. Further, clinicians may find additional tools overwhelming and could misuse them. We derived five design strategies that guide the development of our AI-based CDSS based on the clinicians' needs to support them in their decision-making and to sensitize and guide them.

4 SIT-CARE: Human-centered AI-based support system for ASC screening

In this section, we introduce the components of the SIT-CARE system. First, we describe how the elements of the Front-End User Interface were developed based on the Design Strategies identified in Phase I, see Section 3. Then, we go through the necessary back-end calculations to allow the functionalities in both interaction modes of the SIT-CARE

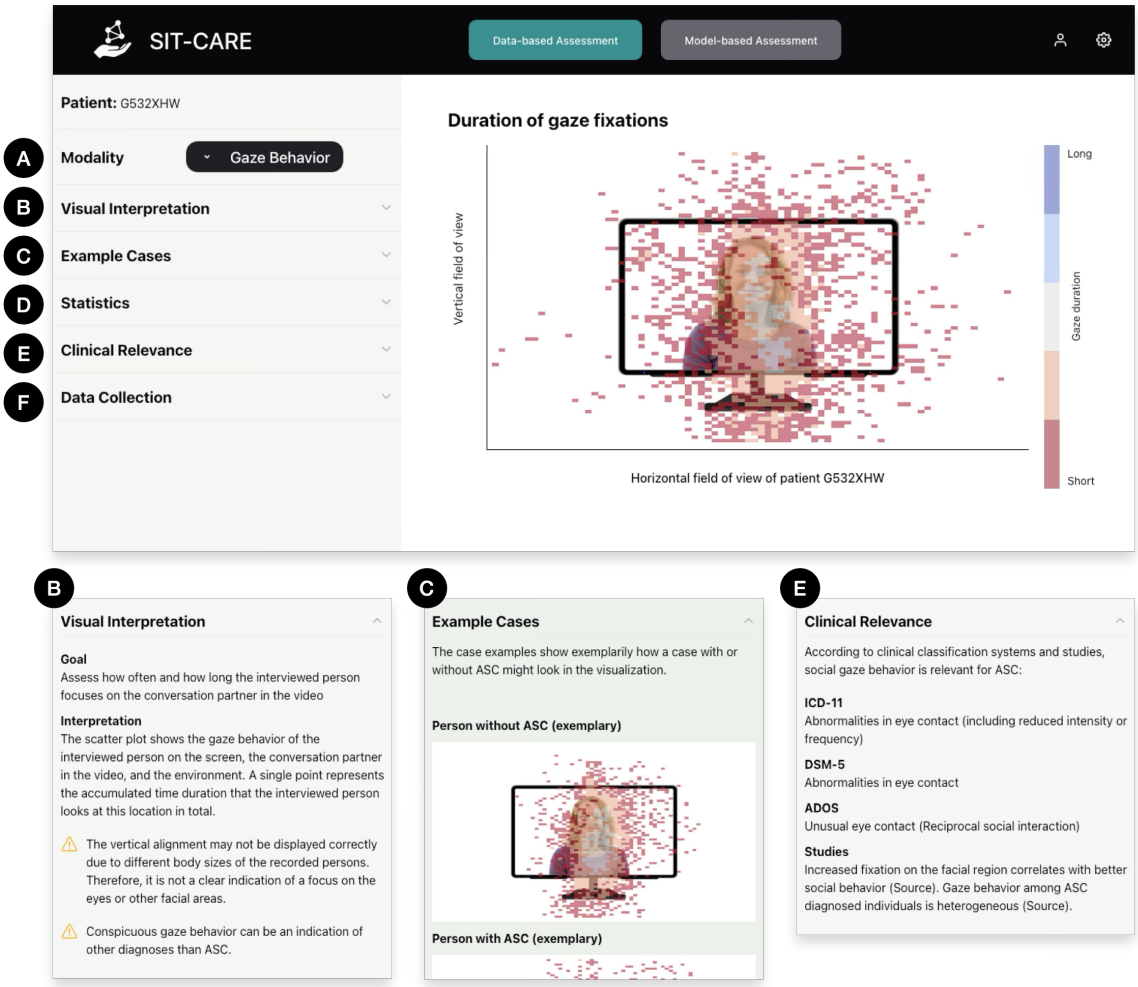

Fig. 1. Here two images are needed: Data- and model-based assessment.

system, Data-based assessment and Model-based assessment. Lastly, we describe the implementation details of the system.

4.1 Front-End User Interface Design

The Front-End User Interface of the SIT-CARE system was developed in an iterative process. We started with a paper prototype to explore initial ideas, which were translated into a functional prototype with dummy visualizations to further improve the system based on gathered feedback from all authors. The final visualizations used in the SIT-CARE system for the non-behavioral patient data were based on identified clinicians' preferences from Phase I and then further iterated by the authors. Textual content of the application, such as the clinical information and warnings, was created based on literature and the domain expertise of the authors, which have a background in psychology, medicine and HCI.

4.1.1 Structure of the user interface. As the findings of phase I show, see Section 3.2, clinicians do not only want an AI recommendation, but want additional support that empowers them in their decision-making process. As a result, the SIT-CARE consists of two main modes. First, DATA-BASED ASSESSMENT supports clinicians in earlier decision-making stages [64, 80, 82] by sensitizing to consider ASC during screening and educating them by providing visualizations and information about non-verbal behavior (hypotheses generation, data gathering); see Figure 1. Second, the MODEL-BASED ASSESSMENT supports intermediate and later decision-making stages, i.e., data gathering and testing hypothesis [64], by providing recommendations and confidence levels for the assessment of non-verbal behavior, which aids in determining whether a full ASC diagnosis procedure is appropriate image autoref. Thus, the SIT-CARE serves as additional support for assessing non-verbal behavior in earlier and later decision-making stages in ASC screening (**Design Strategy 1**). In addition to these two main modes, clinicians enter SIT-CARE via a start screen explaining that only non-verbal behavior is assessed and that the system cannot be used as a standalone. This aligns with **Design Strategy 5** and the human-AI interaction guideline "Make clear what the system can do." [3]. Further, to avoid information overload **(Design Strategy 3)**, we progressively disclosed all described additional information in the sidebar by collapsing all categories in a drop-down list by default [65], which can lead to higher user acceptance in time-constrained environments [31, 76]. In the following, both assessment modes of the system are described.

4.1.2 Data-based assessment mode. The DATA-BASED ASSESSMENT mode of SIT-CARE allows clinicians to explore non-verbal behavioral data of a patient. Following clinicians' advice, we only included modalities relevant in current classification systems (ICD-10, ICD-11, DSM-5), research studies, and the ADOS [52], leading to the inclusion of eye gaze behavior [58], intensity and variability of facial expressions [69], and voice parameters [27] *(Design Strategy 3)*. Providing these partial insights into the training data may increase transparency, which could improve clinicians' understanding of the AI model in the MODEL-BASED ASSESSMENT [41].

For a given patient, the main content in DATA-BASED ASSESSMENT mode shows a visualization of the currently selected modality, i.e., *gaze behavior*, *facial expression* or *voice* (Figure 1 A). See Figure 1 for gaze behavior, where a bi-dimensional histogram plot is used. This visualization was adjusted with clinician feedback to include an illustration of a display and the position of the actress in the SIT, to make the interpretation more intuitive **(Design Strategy 2)**. Line plots were used to depict the variability and intensity of facial expressions over time, which where improved through clinicians' suggestions to include a reference area based on standard deviations from the reference group mean to make it more understandable **(Design Strategy 2)**, see Appendix C. For voice parameters, we focused on pitch variability as this feature was judged to be relevant by the clinicians and is well-known in the scientific literature [6, 45, 56] and research [Anonymized]. Based on clinician feedback, initial box plots were replaced for violin plots to visualize pitch variability with respect to Non-ASC reference distributions **(Design Strategy 2)**, see Appendix C. The line and violin plots were divided into three sections (neutral, joy, disgust) corresponding to the interaction phases of the SIT. Via radio buttons, clinicians could then compare the line and violin plots of the current patient case to different, gender-specific reference groups (**Design Strategy 4**). A hover element provides additional information about the reference groups (**Design Strategy 1**).

To further guide clinicians, a sidebar with additional information on the current modality is integrated; see Figure 1 on the left. Under "Visual Interpretation", see Figure 1 B, information is provided on how to read the current visualization in order to understand and interpret it correctly (*Design Strategy 1*). Following **Design Strategy 4**, prototypical examples [57] of how visualization for each modality could look for an individual with or without ASC are provided, which allows clinicians to compare exemplary case data with the current patient (Figure 1 C). Summary statistics of the

patient for the current modality are presented in a tabular format alongside group-level means and standard deviations of ASC and Non-ASC reference groups (Figure 1 D). To connect to familiar concepts and guide less experienced clinicians (**Design Strategy 1 and 4**), we provide information about the clinical relevance per modality (Figure 1 E).

4.1.3 Model-based assessment mode. Switching to the Model-based assessment mode allows clinicians to inspect the output of the binary AI classification model, which provides a recommendation whether a comprehensive diagnosis of autism should be conducted **(Design Strategy 1)**. Limitations were communicated via warnings below the recommendation. For example, clinicians were reminded that more data was considered than in the data-based assessment and that masking should be considered **(Design Strategy 5)**. To improve interpretability, the recommendation is presented in frequency format [18], indicating the confidence of the AI model and therefore revealing "information about the prediction" [41]. "Information about the model" [41] is provided via the left sidebar, such as insights into the accuracy of the model and a contextualized confusion matrix [3, 63] **(Design Strategy 5)**.

4.2 Data processing and modeling

During the SIT the patient is video- and audio-recorded. From these recordings, non-verbal behavioral features were extracted. All extracted features were aggregated into phase-specific representations (listening vs. speaking) across emotion-eliciting interaction phases (neutral, joy, disgust), enabling targeted analysis of both expressive and receptive behavior. In SIT-CARE, these extracted features are used to generate visualizations and numerical summaries of gaze behavior, facial expressions and voice parameters within the Data-based assessment. Further, for Model-based assessment, we trained AI models using additionally derived non-verbal features for ASC classification beyond the three introduced modalities. In the next section, the back-end calculations are described. Detailed formulas and sampling procedures are provided in the appendix D.

Gaze Behavior: Eye gaze behavior was characterized through geometric transformations of gaze vectors, enabling the computation of metrics such as gaze variability, fixation duration on the display or actress in the video, and frequency of gaze aversion. Gaze angles were transformed into a screen-centered coordinate system to visualize fixation locations relative to the actress. Heat maps were generated by color-coding fixation density into five discrete levels, defined relative to a Non-ASC reference group (see Appendix D). Prototypical examples from individuals of the ASC and the Non-ASC groups were sampled. In addition, the variance of horizontal gaze angles and the percentage of time spent looking at the screen and at the actress's face region were calculated.

Facial Expressions: Facial expressions were extracted using OpenFace 2.2 [9] and quantified via facial action units (AUs). We focused on AU06 (cheek raiser) and AU12 (lip corner puller), representing smiling intensity. For each assessed individual, the mean AU intensity traces were plotted across the interaction using line plots. To allow comparing the current case with Non-ASC individuals, two reference bands were overlaid for comparison: a dark band representing one and a lighter band representing two standard deviations. Representative individuals were selected using a distance-to-median score (Appendix D) and visualized as prototypical examples, with 15% noise added to the traces to ensure anonymity. Further, the mean and variance of smiling intensity (AU06 + AU12), mean and variance of AU04 (brow lowering), and overall mean intensity across all AUs were calculated.

Voice Parameters: Prosodic aspects of speech were extracted using the openSMILE [25] toolkit, including the features pitch, intensity, jitter, shimmer, and harmonic-to-noise ratio. Pitch variability measures the range of fundamental frequency (F0, in semitones relative to 27.5 Hz) between the 2nd and 98th percentiles of voiced segments (Appendix D).

Prototypical example cases corresponded to group medians. Further, pitch variance, mean and variance of voiced segment length, and loudness variability were calculated.

AI Classification Model: Beyond the selected features used in the Data-based assessment, additional multimodal features were used to train a binary classification model (ASC, Non-ASC). For instance, head motion parameters were additionally extracted and processed using the OpenFace library. In total, 1140 features were used to train the late-fusion binary classification model following the approach established in our earlier work [Anonymized]. Thus, each modality was modeled separately using XGBoost (gradient-boosted decision trees). The resulting probability scores were combined using logistic regression with polynomial features (degree 2) to capture non-linear interactions. The model was trained on a dataset of 325 adult participants (168 ASC, 157 Non-ASC; gender-balanced). Performance was evaluated using participant-based Leave-One-Out Cross-Validation (LOOCV [77]) to ensure robust generalization. In that evaluation, the late-fusion model achieved an accuracy of 74%, with a precision of 0.76 and a recall of 0.74, outperforming unimodal approaches. For the present user study, we used this modeling pipeline and retrained the model on the entire dataset of 325 participants. The retrained model was then applied to classify two newly recorded cases that were not part of the training dataset.

4.3 System Implementation

The SIT-CARE system was developed as a typescript-based web application. For the front- and back-end implementation, the *NextJS/React* framework was used, *ChakraUI* served as visual component library. Visualizations of patient modalities were generated in *Python* using `matplotlib` and `seaborn` and then integrated into the web application. While contextual elements such as legends, layouts, and interactions with visualizations were implemented directly in the web application, the visualizations needed to be pre-rendered as due to the high data privacy requirements of working with actual ASC patient data, the underlying data were not allowed to reside directly on the server for dynamic rendering. The application was deployed on institutional servers, and study participants were granted access through password-protected individual links.

5 User Study: Understanding and Supporting AI-assisted Clinical Decision-making

To investigate the influence of SIT-CARE on clinicians' mental models during their decision-making and to answer our second research question, we conducted interviews with clinicians during which they interacted with the SIT-CARE and assessed two patient cases. Per case they made three decisions at three consecutive decision points (DP): after seeing the SIT video (DP1), after exploring the SIT-CARE's data-based assessment (DP2) and lastly, after seeing the model-based assessment (DP3). Our findings summarize what decision paths clinicians took and how their mental models played a role in each decision. Further, clinicians' expressed high interest in using SIT-CARE in their ASC screening or as a learning opportunity.

5.1 Method

5.1.1 Participants, Data Collection and Analysis. We recruited seven clinicians with differing experience in clinical psychology and ASC via purposive sampling [4], see Table 3 for demographics. We conducted an open-ended, semi-structured interviews [44] remotely each lasting around 68 minutes. Approximately 8 hours of audio material was transcribed, which we inductively and deductively coded [48]. To understand the participants' decision-making process, we categorized the paths that the clinicians followed. This allowed us to explore the mental models that led a clinician to

Table 3. Demographics of Participants in our User Study.

P#	Gender	Professional Stage	Experience
P8	Female	Practicing psychotherapist	4 years
P9	Female	Practicing psychotherapist	40 years, experience with ASC
P10	Male	Practicing psychotherapist	12 years
P11	Male	Practicing psychotherapist	30 years, specialized in ASC
P12	Male	Practicing psychotherapist	40 years, specialized in ASC
P13	Female	Practicing psychotherapist	11 years, specialized in ASC
P14	Female	Practicing psychotherapist	>4 years, first experiences with ASC

take a specific path. We analyzed the transcripts by reading them multiple times and derived the initial codes and themes iteratively. The findings were collaboratively structured and refined through iterative feedback and team discussions to ensure consensus. The study was approved by our IRB[4]. All participants consented prior to participation.

5.1.2 Study Procedure. First, we asked demographic and work-related questions, then introduced the SIT-CARE, which aims to support in assessing non-verbal behavior for diagnosing ASC. Next, the clinicians were asked to assess two patient cases, which were chosen by experienced clinicians based on their diagnostic assessments (first case: Non-ASC, second case: ASC) and are considered clear cases. We obtained the patients' informed consent to share their data. For each case, starting with the Non-ASC case, the participants first saw the video recording of the person conducting the SIT. Next, they explored the SIT-CARE's DATA-BASED ASSESSMENT (see Section 4) and then the MODEL-BASED ASSESSMENT. After each step, the clinicians were asked if the patient's non-verbal behavior indicated ASC and to explain their reasoning. Thus, per case the clinicians made three decisions, i.e., video-based decision **(Decision Point 1)**, DATA-BASED decision **(Decision Point 2)**, and MODEL-BASED decision **(Decision Point 3)**. To be able to explore their mental models about the AI (see [30]), participants were asked to predict what the AI may suggest and why at DP1 and DP2. During the task, participant verbalized their cognitive process (think-aloud protocols), which have shown to be useful for exploring decision processes in interactions with CDSS [1, 68, 71]. Additional questions were asked such as about the helpfulness, trustworthiness of the SIT-CARE and their general impression, see Appendix B .

5.2 Findings: Evaluation and integration of the SIT-CARE tool

In this section, we summarize the clinicians' opinions about the SIT-CARE, i.e., the DATA-BASED ASSESSMENT, followed by the feedback about the MODEL-BASED ASSESSMENT, and lastly, the integration and use of the SIT-CARE in practice.

5.2.1 Diverse Expectations of Data-based Assessments. Overall, most participants found the data-based assessment to be helpful and reliable. Especially the gaze visualization and the prototypical examples were liked. However, opinions on the level of detail varied.

The gaze information was described as intuitive and "*really good represented graphically*" (P8). However, the violin plots were described as rather unfamiliar (P8, P14) and the reference area for the line plots could be misunderstood (P11). Participants had different opinions about the type and level of detail of the information Some participants expressed interest in detailed summary statistics (e.g., per SIT phase and gender-specific (P12)), analysis of gestures and what

[4]Approval number and institution anonymized for review.

was said (P8, P11). In contrast, others expressed that the different types of visualizations may already be challenging. P10 stated: "*The less familiar the person is with scientific work, the more difficult it will probably be to understand such graphs*". P13 recommended to focus only on "*one [non-verbal behavior] that is classically presentable, for example, gaze behavior*" and underlined that detailed statistics would not be helpful. In addition, participants mentioned that they need an "*introductory module to sort out this mass of information and data [...] so that [they] know what to do with all the information*" (P13) and more information on "*why [some behavior] is somehow typical for autism*" (P14). To guide clinicians, P10 recommended to highlight in the visualizations "*what you would expect to stand out from a normal, [Non ASC] control group [...]*". In addition, the prototypical examples were received very well, and it was recommended to include more examples depicting a range of possible ASC behavior (P10, P11).

5.2.2 Feedback on the Model-based Assessment. Summarized, participants had mixed opinions on the MODEL-BASED ASSESSMENT, with some considering it more than the DATA-BASED ASSESSMENT and others disregarding it completely.

Two participants reported that they "*ignored*" (P9) the MODEL-BASED ASSESSMENT and "*found [one case] to be misjudged*" (P11). Others emphasized the advantage of THE MODEL-BASED ASSESSMENT, which is "*time-efficient*" (P13) and provides "*additional information that the normal human brain cannot fully process*" (P13) without help, "*while the data-based assessment only exemplarily represents three areas*" (P10). P12 described that the MODEL-BASED ASSESSMENT interrupted their typical decision-making process: "*I was deviating from my usual assessment, and I felt uncertain. [...] There's the value, and what does that actually mean, and does that align with my impression or the questionnaire values? So that irritated me a little at first, which isn't such a bad thing, to recalibrate yourself in diagnostics as well. It creates a bit of awareness to take a closer look at yourself.*" Some participants preferred such "*a clear value*" (P13), others warned that a number as "*bit tempting*" which may lead to over-reliance ignoring aspects like masking (P12). This risk was partially mitigated by integrating a confidence matrix and warnings, which were noted as "*very transparent, good, and important*" (P10). Further adjustments were recommended such as providing more information about the AI model (P9), considering gender (P12) and what was said (P11). Participants also discussed whether they want to be able to calibrate the AI model's sensitivity and specificity, but P13 argued, that this would hinder comparability.

5.2.3 Between Supporting ASC Screening and Raising Awareness. Overall, participants expressed high interest in the SIT-CARE as an ASC screening support and as a learning opportunity.

All participants expressed interest in the SIT-CARE tool and indicated that they would use it "*Yes, immediately. Yes, absolutely.*" (P12), and were open for AI-based support "*if there's a trustworthy organization behind it*" (P12). Two purposes of the SIT-CARE were highlighted: First, to use it in everyday practice to support ASC diagnostics and screening as the "*autism screening instruments that currently exist [...] are simply so outdated that they don't give much anymore*" (P13), and second, to raise awareness and educate oneself about the topic (P10, P12, P13). P12 stated "*that it can definitely help raise awareness of a possible ASC diagnosis that I would otherwise overlook*". Further, all participants expressed a need for guidance in the form of additional information and an onboarding process. For example, P13 suggested "*a short video would be helpful to explain [the data-based assessment]*" (P13). Moreover, in both studies, participants expressed an interest in expanding the SIT-CARE beyond ASC. P2 were unsure if it is possible to differentiate between disorders, but "*[they] think there are probably specific characteristics that apply to specific disorders*" and that "*whether that's really specific would be totally cool to find out.*"

5.3 Findings: Decision-making Process and Users' Mental Models

In the following, we describe decision paths the clinicians followed during their decision-making process per case. Seven clinicians assessed each of the two cases at three decision points, thus, making 42 decisions. Per assessment, the clinicians took a specific path, which is characterized by what decision was made at each decision point, and whether the decision was updated after interacting with SIT-CARE, see Figure 2. We summarize these decision paths and explore clinicians' mental models during this process.

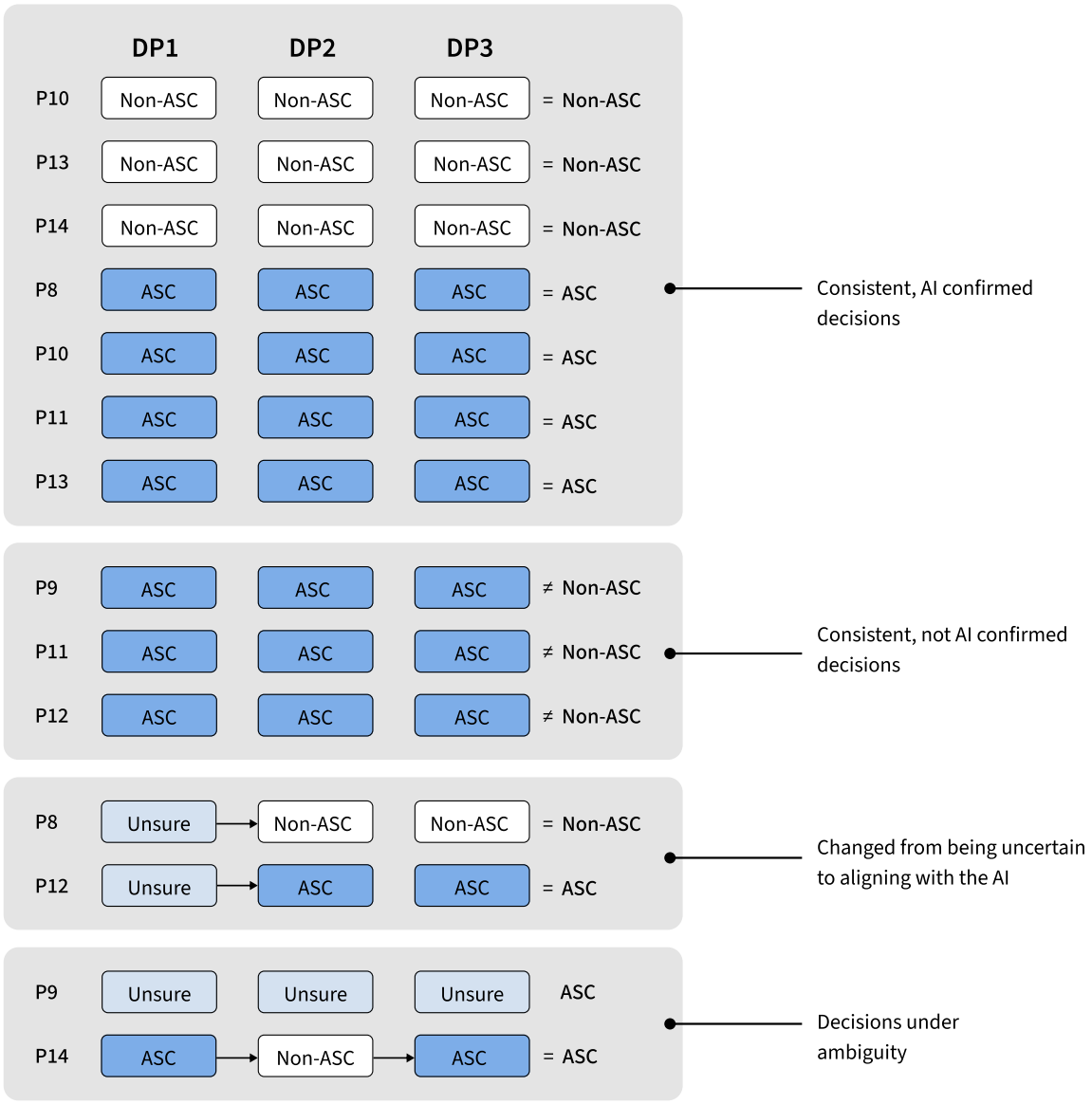

Fig. 2. TODO Decision changes of the clinicians over the three decision points, i.e., video-based assessment (DP1), data-based assessment (DP2), and model-based assessment (DP3).

5.3.1 Different Reasoning can lead to the Same Decision. In seven of the fourteen decisions to be made, the clinicians did not deviate from their initial decision, which was then confirmed in the MODEL-BASED ASSESSMENT. Their decision paths were similar, but they considered SIT-CARE information differently.

Three of the participants made **consistent, AI confirmed decisions**. They considered the circumstances, and described them as "*artificial*" (P10) and "*in light of this situation, [they] found the emotional expression to be relatively pronounced*" (P13). However, their reasoning differed. P14 interacted extensively with the data-based assessment, read

all materials in detail, considered the prototypical examples and summary statistics: "*I find the data helpful because it captures much more than I would notice in person*". P8 appreciated the clinical information about masking behavior as they are less experienced, and expressed: "*It kind of confirms me, because I was totally unsure*". P13, however, only considered their own and the model-based assessment in their final decision and stated "*the way the AI collects its data is so different from my way of perceiving this information, so far apart, that I wouldn't even try to compare it.*" Regarding the second case (ASC), only the participants persisting on their initial assessment which was confirmed by the AI highlighted the person's "*stimming behavior*" (P11), "*odd hand movement*" (P13) and "*stereotypical movements*" (P8). For all of the described decisions above, the participants predicted the AI to come to the same conclusion and confirmed their assessment. For example, P8 very confidently predicted a high confidence of the AI, and was surprised by the real confidence value.

5.3.2 Perceived Conflict between the Data-based and Model-based Assessment. Three participants rated the first case (Non-ASC) as rather typical for ASC at each decision point, disagreeing with the MODEL-BASED ASSESSMENT. Thus, they made **consistent, not AI confirmed decisions**. Although the DATA-BASED ASSESSMENT prompted reflection, the video-based impressions were emphasized more.

After seeing the video (P1), they considered not only non-verbal behavior, but also what was said. P9 specified: "*I don't think others [without ASC] would quickly grasp, the neutral flavor that has been extracted [..], I found that to be typically autistic.*" After reviewing the data-based assessment, one participant was still confident in their decision (P11), others mentioned "*a few aspects, like this gaze into the eyes, [as] rather untypical for autism*" (P12). Still, they assumed SIT-CARE would confirm them and were surprised by the MODEL-BASED ASSESSMENT. P11 stated "*that irritates [me] because [...] that was relatively autism-typical.*". P12 reiterated that in this case, the person may mask behavior. All three participants considered the DATA-BASED ASSESSMENT (P9, P12) and their video-based assessment as most insightful, as they "*trust [themselves] the most*" (P9).

5.3.3 From Indecision to Decision. For two decisions, participants **changed from being uncertain to aligning with the AI** after exploring the DATA-BASED ASSESSMENT, which helped them find indicators that led to a more confident assessment.

Based on the video for the second case (ASC) "*[they] didn't find it clearly typical [...] as there was too much there.*" (P12). Another participant noticed contradicting indications, as the person smiled appropriately but also seemed "*petrified*" (P8). Both were unsure of what the AI may recommend, stating "*they found [the video] too short to assess*" (P8). After exploring the DATA-BASED ASSESSMENT, their decisions changed as they identified guiding indicators. P8 now assumed it is "*a person without autism, because they actually show a lot more in their facial expression or even matching facial expressions to what's happening.*", but refrained from predicting the AI's recommendation. P12 said the DATA-BASED ASSESSMENT clarified the position of the display and that "*with the individual data, the scales tipped slightly in the direction of autism, so I would have recommended further diagnostics.*" The model-based assessment manifested their assessments. They described "*found it again a bit of a key that strengthened [their] tendencies*" (P12).

5.3.4 Decisions under Ambiguity. Two decision paths were shaped by **ambiguity**, showing the risk of indecision and the lack of comprehensive comparisons to other diagnoses.

For the second case (ASC), P9 was indecisive at all three decision points and after seeing the video (DP1) explained that the person appeared "*not particularly lively, but that can also be due to depression*" and reiterated their surprise about the AI's first recommendation. After seeing the model-based assessment (DP3) they underlined: "For me, depression is

simply in the foreground." In contrast, P14 changed their decision twice, first leaning towards ASC based on gaze and facial expressions (DP 1), then changing their mind (DP2) because "*if you look at the numbers, it will be more untypical [for ASC] from the facial expression and also from the voice.*" At each DP, they assumed the AI agreed with their current assessment. They revised their decision back to ASC after seeing the model-based assessment (DP3).

6 Discussion

We designed SIT-CARE based on a formative study of clinicians' workflows and needs. It presents nonverbal behavioral features from SIT videos in two modes. The first mode is a DATA-BASED ASSESSMENT that provides various visualizations and additional information, such as prototypical examples and summary statistics. The second mode is a MODEL-BASED ASSESSMENT that recommends whether to perform comprehensive ASC diagnostics. In a user study, we examined how SIT-CARE influences clinicians' mental models and decision-making processes.

6.1 Effects of Data- and Model-based Support on Clinicians' Mental Models

Two real cases (ASC, Non-ASC) each were assessed by seven clinicians, which made a total of 42 decisions within 14 decision paths.

Our study revealed that, although heterogeneous, clinicians' decision-making processes can be divided into four groups.

The first group, "Different Reasoning Can Lead to the Same Decision" accounts for half of all decision paths. In this group, the clinicians' initial judgments were unchanged and confirmed by the MODEL-BASED ASSESSMENT. However, the clinicians weighed the information provided by SIT-CARE differently, and underline the need to research clinicians' mental models as from the decision alone their reasoning can not be implied conclusively. Interestingly, participants commonly anticipated concurrence from AI across all decision paths. However, this does not mean that clinicians accurately calibrated their confidence in model performance. Rather, it underscores the need for improved communication and training regarding confidence levels.

The second group, "Perceived Conflict Between Data-Based and Model-Based Assessments" has three decision paths in which clinicians consistently misjudged the non-ASC case as ASC at every decision point, thus also diverged from the MODEL-BASED ASSESSMENT. Although the DATA-BASED ASSESSMENT prompted reflection, it did not override video-based impressions. This persistent disagreement between humans and AI may be due to different lines of reasoning. These under-reliant clinicians prioritized interactive video cues and experiential heuristics. Providing guidance, such as through counterfactual explanations, could help clinicians determine when to defer to or stand by their judgment in ambiguous cases.

The third group, "From Indecision to Decision" consists of two decision paths. After using the DATA-BASED ASSESSMENT, clinicians shift from uncertainty to alignment with the AI. Initially, video impressions yielded ambiguous and conflicting cues, so participants could not predict the AI's recommendation. However, the DATA-BASED ASSESSMENT provided guiding indicators that resolved the ambiguity and increased confidence. The subsequent MODEL-BASED ASSESSMENT confirmed the clinicians' revised judgments. Therefore, when initial evidence is inconclusive, data-based assessments can effectively guide sensemaking without overriding clinicians' autonomy. Progressive disclosure of data could facilitate this transition.

The final group, "Decisions Under Ambiguity" shows how ambiguity can lead to indecision, especially when comparisons to other possible diagnoses are not given. This underscores the importance of tools that contrast ASC indicators with alternative diagnoses, such as depression, in order to avoid anchoring. Displays showing how and

why one's assessment changes over time may help clinicians maintain consistency and calibrate their reliance during ambiguous cases.

6.2 Standardized Screening Support and Learning Opportunity

In the user study, clinicians described two main use cases for SIT-CARE: (1) supporting the diagnostic process across early to late stages, and (2) training less experienced clinicians to assess the non-verbal behavior associated with autism spectrum conditions. For supporting the diagnostic process, the SIT-CARE could support in the screening stages, i.e., before detailed diagnostics, by determining which individuals should be referred to specialized clinicians, replacing or complementing questionnaires of uncertain reliability (ref AQ). Since the SIT is a web-based tool that only takes a few minutes to complete and can be administered on a laptop with a standard webcam at home, it has strong scalability potential. In prior work [Anonymized], we demonstrated that the SIT yields comparable accuracy in home environments relative to standardized laboratory setting. At a later diagnostic stage, especially for ambiguous cases, the SIT-CARE can provide objective information about non-verbal behavior from a standardized paradigm to support clinical decision making. In both early and late use cases, clinicians can interact with the system's statistics, visualizations and model recommendations independently of the client's location, supporting integration into telehealth workflows. The education potential of SIT-CARE is highly relevant given common misconception of ASC among general practitioners (ref) and the shortage of specialized clinicians (ref). By enabling exploration of the current case and prototypical cases, SIT-CARE helps young clinicians develop diagnostic sensibility for assessing non-verbal behaviors in autism. Notably, SIT-CARE includes male and female prototypical cases, raising awareness of the female autism phenotype, which is often overlooked (ref). Whether the clinicians view the SIT-CARE as a screening tool, a means to enrich the diagnostic process, or an educational support, appears to depend on the expertise level of potential user and the ambiguity of a diagnostic case.

- discuss possible causes for different perspectives: different needs and purposes - people would like it as a Screening tool, potentially also as a learning tool - allgemeine Vorteile des SIT: standardized task, scalable - wie zukünftig eingesetzt - potentiale
- relate to guidance research (our and the other articles discuss in the previous research) - Claudia
- Connection Section: 5.2 Findings: Evaluation and integration of the SIT-CARE tool 5.2.3 Between supporting screening and raising awareness,
- Anbinden an Overreliance, zu sehr überzeugend (Zb nach Data based Meinung geändert, und dann wieder zurück geändert)

6.3 Ethical Considerations in using AI for the Diagnosis Process of ASC and Beyond

Our research focused on creating an interactive user interface for SIT. However, applying the prototype in a clinical setting raises a number of ethical concerns.

Due to waitlists and workforce shortages, institutions may use AI outputs as de facto diagnoses, which contradicts the intended role of SIT-CARE as a decision support tool. Thus, safeguards are needed, such as scope-of-use messaging ("recommendation to consider further diagnostics" rather than "diagnosis") []. Additionally, model confidence, data quality, and the evidence basis should be provided in the form of model cards and data sheets []. Furthermore, clear clinical AI usage guidelines are needed (see []). For example, clinicians could be required to document and sign off on the supporting evidence before taking downstream actions.

Currently, diagnosis relies on standardized instruments and expert judgment acquired over years. AI decision support can shift epistemic authority and alter tacit skill formation []. Evidence from related fields suggests that diagnostic accuracy may decrease with AI assistance [], raising concerns about de-skilling, especially among less experienced clinicians who may become overly dependent on recommendations. This underscores the importance of positioning SIT-CARE as a complement to, rather than a replacement for, comprehensive assessment. For example, incorporating supervised onboarding and periodic skills assessments can ensure appropriate usage and monitor whether clinicians' independent judgment improves over time [].

Furthermore, providing pedagogical resources, such as prototypical examples across subgroups or counterfactuals, can encourage active learning [].

7 Limitations and Future Work

Our work has several limitations. First, we only involved seven clinicians in our formative study and newly recruited seven clinicians in our user study, a sample size similar to prior domain-specific qualitative work (e.g., [13, 17, 80, 82]). To receive professional feedback, most of our recruited clinicians several years of experience and with ASC. Future work is needed to specifically investigate the generalizability on other target groups such as inexperienced clinicians.

Second, our underlying AI model of SIT-CARE was trained on data of adult participants [Anonymized] with a focus on ASC. This data limitation may affect the generalizability and validity of the SIT-CARE due to three aspects: Both, the individuals of the SIT data and the clinicians in our study were of one country, which may limit its generalizability, as nonverbal behavior is strongly influenced by culture [39, 47]. Also, in practice, clinicians typically consider multiple potential diagnoses at once. Other diagnoses where research indicated that non-verbal behavior may deviate from the norm would complement the diagnostic workflow better (e.g., eating disorders [16, 50] and Attention Deficit Hyperactivity Disorder (ADHD) [Source]). Further, the ground truth for training the SIT-CARE's AI model is based on diagnoses made after a full diagnostic procedure, which inherently involves clinical subjectivity. However, training AI models on large-scale SIT datasets could enable the identification of data-driven phenotypes[5] which could become part of the SIT-CARE and thus change the diagnostic process. Future work should focus on increasing the dataset considering culture, differential diagnosis and explore to date unknown ASC markers to improve SIT-CARE's applicability.

Third, using an interview format led to an insightful, detailed understanding of clinicians' decision-making process and their mental models of the AI, data and case. However, large scale quantitative research with clinicians via cluster sampling is needed, to assess the team performance, diagnostic accuracy and to explore the decision paths under real-life circumstances. Thus, paving the way to understand the decision paths and mental models of experts enough to then be able to get an insight into how less experienced clinicians could adjust their decision-making, for example, by providing guidance how information should be weighted.

Lastly, our study focused on a first implementation, future work is needed to deploy the SIT-CARE and make it available to practicing psychotherapists. However, to do so extensive onboarding including detailed information on how to consider specific non-verbal behavior in their reasoning is required. For example, with video tutorials or including the SIT-CARE in further education on ASC for clinicians.

[5] A person's phenotype is the expression of their genotype and can be observed through an individual's appearance, signs, or symptoms of a disease [78].

8 Conclusion

Placeholder for wordcount: Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et.

9 Statement of Use of LLMs

For this paper, LLMs were not used beyond editing our own text.

References

[1] Benjamin Abdel-Karim, Nicolas Pfeuffer, K. Valerie Carl, , and Oliver Hinz. 2023. How AI-Based Systems Can Induce Reflections: The Case of AI-Augmented Diagnostic Work. *MIS Quarterly* 47, 4 (Dec. 2023), 1395–1424. doi:10.25300/MISQ/2022/16773

[2] Carrie Allison, Bonnie Auyeung, and Simon Baron-Cohen. 2012. Toward Brief "Red Flags" for Autism Screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist for Autism in Toddlers in 1,000 Cases and 3,000 Controls [Corrected]. *Journal of the American Academy of Child and Adolescent Psychiatry* 51, 2 (Feb. 2012), 202–212.e7. doi:10.1016/j.jaac.2011.11.003

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–13. doi:10.1145/3290605.3300233

[4] Chittaranjan Andrade. 2021. The Inconvenient Truth About Convenience and Purposive Samples. *Indian Journal of Psychological Medicine* 43, 1 (Jan. 2021), 86–88. doi:10.1177/0253717620977000

[5] K. L. Ashwood, N. Gillan, J. Horder, H. Hayward, E. Woodhouse, F. S. McEwen, J. Findon, H. Eklund, D. Spain, C. E. Wilson, T. Cadman, S. Young, V. Stoencheva, C. M. Murphy, D. Robertson, T. Charman, P. Bolton, K. Glaser, P. Asherson, E. Simonoff, and D. G. Murphy. 2016. Predicting the Diagnosis of Autism in Adults Using the Autism-Spectrum Quotient (AQ) Questionnaire. *Psychological Medicine* 46, 12 (Sept. 2016), 2595–2604. doi:10.1017/S0033291716001082

[6] American Psychiatric Association. 2013/2013. *Diagnostic and Statistical Manual of Mental Disorders : DSM-5™*. (5th edition. ed.). American Psychiatric Publishing, a division of American Psychiatric Association, Washington, DC ;.

[7] Autism and Developmental Disabilities Monitoring Network Surveillance Year 2008 Principal Investigators. 2012. Prevalence of autism spectrum disorders–autism and developmental disabilities monitoring network, 14 sites, United States, 2008. *Morbidity and Mortality Weekly Report: Surveillance Summaries* 61, 3 (2012), 1–19.

[8] Anne Kathrine Petersen Bach, Trine Munch Nørgaard, Jens Christian Brok, and Niels Van Berkel. 2023. "If I Had All the Time in the World": Ophthalmologists' Perceptions of Anchoring Bias Mitigation in Clinical AI Support. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–14. doi:10.1145/3544548.3581513

[9] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (2018-05). 59–66.

[10] 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. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), 2–11. doi:10.1609/hcomp.v7i1.5285

[11] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58 (June 2020), 82–115. doi:10.1016/j.inffus.2019.12.012

[12] Sotirios Batsakis, Marios Adamou, Ilias Tachmazidis, Sarah Jones, Sofya Titarenko, Grigoris Antoniou, and Thanasis Kehagias. 2022. Data-Driven Decision Support for Adult Autism Diagnosis Using Machine Learning. *Digital* 2, 2 (June 2022), 224–243. doi:10.3390/digital2020014

[13] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. doi:10.1145/3313831.3376718

[14] Kellen Briot, Adrien Pizano, Manuel Bouvard, and Anouck Amestoy. 2021. New Technologies as Promising Tools for Assessing Facial Emotion Expressions Impairments in ASD: A Systematic Review. *Frontiers in Psychiatry* 12 (2021), 634756. doi:10.3389/fpsyt.2021.634756

[15] Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong. 2023. Improving Human-AI Collaboration With Descriptions of AI Behavior. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–21. doi:10.1145/3579612

[16] H. Pinar Caglar-Nazali, Freya Corfield, Valentina Cardi, Suman Ambwani, Jenni Leppanen, Olaolu Olabintan, Stephanie Deriziotis, Alexandra Hadjimichalis, Pasquale Scognamiglio, Ertimiss Eshkevari, Nadia Micali, and Janet Treasure. 2014. A systematic review and meta-analysis of 'Systems for Social Processes' in eating disorders. *Neuroscience & Biobehavioral Reviews* 42 (May 2014), 55–92. doi:10.1016/j.neubiorev.2013.12.002

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

[18] Shiye Cao, Anqi Liu, and Chien-Ming Huang. 2024. Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 1–32. doi:10.1145/3637318

[19] Tara Capel and Margot Brereton. 2023. What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–23. doi:10.1145/3544548.3580959

[20] Nadire Cavus, Abdulmalik A. Lawan, Zurki Ibrahim, Abdullahi Dahiru, Sadiya Tahir, Usama Ishaq Abdulrazak, and Adamu Hussaini. 2021. A Systematic Literature Review on the Application of Machine-Learning Models in Behavioral Assessment of Autism Spectrum Disorder. *Journal of Personalized Medicine* 11, 4 (April 2021), 299. doi:10.3390/jpm11040299

[21] 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. doi:10.1145/3290605.3300789

[22] Caitlin M. Conner, Ryan D. Cramer, and John J. McGonigle. 2019. Examining the Diagnostic Validity of Autism Measures Among Adults in an Outpatient Clinic Sample. *Autism in Adulthood: Challenges and Management* 1, 1 (March 2019), 60–68. doi:10.1089/aut.2018.0023

[23] Conor James Davidson, Alwyn Kam, Frances Needham, and Alison Jane Stansfield. 2015. No exclusions – developing an autism diagnostic service for adults irrespective of intellectual ability. *Advances in Autism* 1, 2 (Oct. 2015), 66–78. doi:10.1108/AIA-08-2015-0010

[24] Hanna Drimalla, Tobias Scheffer, Niels Landwehr, Irina Baskow, Stefan Roepke, Behnoush Behnia, and Isabel Dziobek. 2020. Towards the automatic detection of social biomarkers in autism spectrum disorder: introducing the simulated interaction task (SIT). *npj Digital Medicine* 3, 1 (Feb. 2020), 25. doi:10.1038/s41746-020-0227-5

[25] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia* (New York, NY, USA, 2010-10-25) *(MM '10)*. Association for Computing Machinery, 1459–1462.

[26] Laura Fusar-Poli, Natascia Brondino, Pierluigi Politi, and Eugenio Aguglia. 2022. Missed diagnoses and misdiagnoses of adults with autism spectrum disorder. *European Archives of Psychiatry and Clinical Neuroscience* 272, 2 (March 2022), 187–198. doi:10.1007/s00406-020-01189-w

[27] Riccardo Fusaroli, Anna Lambrechts, Dan Bang, Dermot M. Bowler, and Sebastian B. Gaigg. 2017. "Is voice a marker for Autism spectrum disorder? A systematic review and meta-analysis". *Autism Research* 10, 3 (March 2017), 384–407. doi:10.1002/aur.1678 Publisher: Wiley.

[28] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. doi:10.1145/3313831.3376316

[29] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. https://www.researchgate.net/publication/352882174_Human-AI_Complementarity_in_Hybrid_Intelligence_Systems_A_Structured_Literature_Review

[30] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (Feb. 2023), 1096257. doi:10.3389/fcomp.2023.1096257

[31] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. doi:10.1145/3411764.3445385

[32] Sarah L Jones, Maria Johnson, Bronwen Alty, and Marios Adamou. 2021. The Effectiveness of RAADS-R as a Screening Tool for Adult ASD Populations. *Autism Research and Treatment* 2021, 1 (2021), 9974791. doi:10.1155/2021/9974791

[33] Shahad Sabbar Joudar, A. S. Albahri, Rula A. Hamid, Idrees A. Zahid, M. E. Alqaysi, O. S. Albahri, and A. H. Alamoodi. 2023. Artificial intelligence-based approaches for improving the diagnosis, triage, and prioritization of autism spectrum disorder: a systematic review of current trends and open issues. *Artificial Intelligence Review* 56, S1 (Oct. 2023), 53–117. doi:10.1007/s10462-023-10536-x

[34] Markelle Kelly, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2023. Capturing Humans' Mental Models of AI: An Item Response Theory Approach. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1723–1734. doi:10.1145/3593013.3594111

[35] Na Yeon Kim, Junfeng He, Qianying Wu, Na Dai, Kai Kohlhoff, Jasmin Turner, Lynn K. Paul, Daniel P. Kennedy, Ralph Adolphs, and Vidhya Navalpakkam. 2024. Smartphone-Based Gaze Estimation for in-Home Autism Research. *Autism Research* 17, 6 (2024), 1140–1148. doi:10.1002/aur.3140

[36] Anne-Kathrin Kleine, Eesha Kokje, Pia Hummelsberger, Eva Lermer, Insa Schaffernak, and Susanne Gaube. 2025. AI-enabled clinical decision support tools for mental healthcare: A product review. *Artificial Intelligence in Medicine* 160 (Feb. 2025), 103052. doi:10.1016/j.artmed.2024.103052

[37] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell Me More?: The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Austin Texas USA, 1–10. doi:10.1145/2207676.2207678

[38] Anil Kumar and Umesh Chandra Jaiswal. 2025. Predicting Autism Spectrum Disorder in Adults Through Facial Image Analysis: A Multi-CNN with BiLSTM Model. *SN Comput. Sci.* 6, 3 (March 2025). doi:10.1007/s42979-025-03783-y

[39] Marianne LaFrance and Clara Mayo. 1978. Cultural aspects of nonverbal communication. *International Journal of Intercultural Relations* 2, 1 (1978), 71–89. Publisher: Elsevier.

[40] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–18. doi:10.1145/3491102.3501999

[41] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1369–1385. doi:10.1145/3593013.3594087

[42] Silke Lipinski, Elisabeth S. Blanke, Ulrike Suenkel, and Isabel Dziobek. 2019. Outpatient Psychotherapy for Adults with High-Functioning Autism Spectrum Condition: Utilization, Treatment Satisfaction, and Preferred Modifications. *Journal of Autism and Developmental Disorders* 49, 3 (March 2019), 1154–1168. doi:10.1007/s10803-018-3797-1

[43] Silke Lipinski, Katharina Boegl, Elisabeth S Blanke, Ulrike Suenkel, and Isabel Dziobek. 2022. A blind spot in mental healthcare? Psychotherapists lack education and expertise for the support of adults on the autism spectrum. *Autism* 26, 6 (Aug. 2022), 1509–1521. doi:10.1177/13623613211057973

[44] Robyn Longhurst. 2003. Semi-structured interviews and focus groups. *Key methods in geography* 3, 2 (2003), 143–156.

[45] Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H Cook, Bennett L Leventhal, Pamela C DiLavore, Andrew Pickles, and Michael Rutter. [n. d.]. The Autism Diagnostic Observation Schedule–Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism. ([n. d.]).

[46] Catherine Lord, Michael Rutter, and Ann Le Couteur. 1994. Autism Diagnostic Interview—Revised: A Revised Version of a Diagnostic Interview for Caregivers of Individuals with Possible Pervasive Developmental Disorders. *Journal of Autism and Developmental Disorders* 24, 5 (1994), 659–685. doi:10.1007/BF02172145

[47] Áine Lorié, Diego A. Reinero, Margot Phillips, Linda Zhang, and Helen Riess. 2017. Culture and nonverbal expressions of empathy in clinical settings: A systematic review. *Patient Education and Counseling* 100, 3 (March 2017), 411–424. doi:10.1016/j.pec.2016.09.018

[48] Philipp Mayring. 2014. *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. Klagenfurt, Austria.

[49] Goldie A McQuaid, Nancy Raitano Lee, and Gregory L Wallace. 2022. Camouflaging in autism spectrum disorder: Examining the roles of sex, gender identity, and diagnostic timing. *Autism* 26, 2 (Feb. 2022), 552–559. doi:10.1177/13623613211042131

[50] Alessio Maria Monteleone, Giammarco Cascino, Valeria Ruzzi, Niccolò Marafioti, Luigi Marone, Roberta Croce Nanni, and Alfonso Troisi. 2022. Non-verbal social communication in individuals with eating disorders: an ethological analysis in experimental setting. *Eating and Weight Disorders - Studies on Anorexia, Bulimia and Obesity* 27, 8 (July 2022), 3125–3133. doi:10.1007/s40519-022-01442-2

[51] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2024. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 1–39. doi:10.1145/3641022

[52] Dan Nakamura, Yoichi Hanawa, Shizuka Seki, Misato Yamauchi, Yuriko Iwami, Yuta Nagatsuka, Hirohisa Suzuki, Keisuke Aoyagi, Wakaho Hayashi, Takeshi Otowa, and Akira Iwanami. 2024. Predictive model using autism diagnostic observation schedule, second edition for differential diagnosis between schizophrenia and autism spectrum disorder. *Frontiers in Psychiatry* 15 (Dec. 2024). doi:10.3389/fpsyt.2024.1493158 Publisher: Frontiers Media SA.

[53] Ranjani Narayanan, Sarah E. Walsh, and Karen M. Feigh. 2023. Development of Mental Models in Decision-Making Tasks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 67, 1 (Sept. 2023), 767–773. doi:10.1177/21695067231192195

[54] Donald A Norman. 1988. *The Psychology of Everyday Things*. Basic books.

[55] Donald A Norman. 2014. Some Observations on Mental Models. In *Mental Models*. Psychology Press, London, 15–22.

[56] World Health Organization. 2004. *ICD-10 : international statistical classification of diseases and related health problems : tenth revision*. World Health Organization.

[57] Antonin Poché, Lucas Hervier, and Mohamed-Chafik Bakkay. 2023. Natural example-based explainability: a survey. In *World conference on eXplainable artificial intelligence*. Springer, 24–47.

[58] Jacqueline A. Riddiford, Peter G. Enticott, Alex Lavale, and Caroline Gurvich. 2022. Gaze and social functioning associations in autism spectrum disorder: A systematic review and meta-analysis. *Autism Research* 15, 8 (Aug. 2022), 1380–1446. doi:10.1002/aur.2729 Publisher: Wiley.

[59] Riva Ariella Ritvo, Edward R. Ritvo, Donald Guthrie, Max J. Ritvo, Demetra H. Hufnagel, William McMahon, Bruce Tonge, David Mataix-Cols, Amita Jassi, Tony Attwood, and Johann Eloff. 2011. The Ritvo Autism Asperger Diagnostic Scale-Revised (RAADS-R): A Scale to Assist the Diagnosis of Autism Spectrum Disorder in Adults: An International Validation Study. *Journal of Autism and Developmental Disorders* 41, 8 (2011), 1076–1089. doi:10.1007/s10803-010-1133-5

[60] Emely Rosbach, Jonas Ammeling, Sebastian Krügel, Angelika Kießig, Alexis Fritz, Jonathan Ganz, Chloé Puget, Taryn Donovan, Andrea Klang, Maximilian C. Köller, Pompei Bolfa, Marco Tecilla, Daniela Denk, Matti Kiupel, Georgios Paraschou, Mun Keong Kok, Alexander F. H. Haake,

Ronald R. De Krijger, Andreas F.-P. Sonnen, Tanit Kasantikul, Gerry M. Dorrestein, Rebecca C. Smedley, Nikolas Stathonikos, Matthias Uhl, Christof A. Bertram, Andreas Riener, and Marc Aubreville. 2025. "When Two Wrongs Don't Make a Right" - Examining Confirmation Bias and the Role of Time Pressure During Human-AI Collaboration in Computational Pathology. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–18. doi:10.1145/3706598.3713319

[61] William Saakyan, Matthias Norden, Lola Herrmann, Simon Kirsch, Muyu Lin, Simon Guendelman, Isabel Dziobek, and Hanna Drimalla. 2023. On Scalable and Interpretable Autism Detection from Social Interaction Behavior. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–8. doi:10.1109/ACII59096.2023.10388157

[62] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.

[63] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–22. doi:10.1145/3415224 Publisher: Association for Computing Machinery (ACM).

[64] Harold C Sox, Michael C Higgins, Douglas K Owens, and Gillian Sanders Schmidler. 2024. *Medical decision making*. John Wiley & Sons.

[65] 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. doi:10.1145/3374218

[66] Christine K. Syriopoulou-Delli. 2025. Advances in Autism Spectrum Disorder (ASD) Diagnostics: From Theoretical Frameworks to AI-Driven Innovations. *Electronics* 14, 5 (Jan. 2025), 951. doi:10.3390/electronics14050951

[67] Fadi Thabtah. 2019. An Accessible and Efficient Autism Screening Method for Behavioural Data and Predictive Analyses. *Health Informatics Journal* 25, 4 (Dec. 2019), 1739–1755. doi:10.1177/1460458218796636

[68] Peter Todd and Izak Benbasat. 1987. Process Tracing Methods in Decision Support Systems Research: Exploring the Black Box. *MIS Quarterly* 11, 4 (Dec. 1987), 493. doi:10.2307/248979

[69] Dominic A. Trevisan, Maureen Hoskyn, and Elina Birmingham. 2018. Facial Expression Production in Autism: A Meta-Analysis. *Autism Research* 11, 12 (Dec. 2018), 1586–1601. doi:10.1002/aur.2037 Publisher: Wiley.

[70] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8, 12 (Oct. 2024), 2293–2303. doi:10.1038/s41562-024-02024-1

[71] Nicholas P. Vitalari. 1985. Knowledge as a Basis for Expertise in Systems Analysis: An Empirical Study. *MIS Quarterly* 9, 3 (Sept. 1985), 221. doi:10.2307/248950

[72] K. Vogeley, J. C. Kirchner, A. Gawronski, L. Tebartz Van Elst, and I. Dziobek. 2013. Toward the development of a supported employment program for individuals with high-functioning autism in Germany. *European Archives of Psychiatry and Clinical Neuroscience* 263, S2 (Nov. 2013), 197–203. doi:10.1007/s00406-013-0455-7

[73] Fred R Volkmar, Marc Woodbury-Smith, S Macari, RA Oien, J McPartland, and D Stavropoulos. 2022. Diagnostic issues and complexities in autism and related conditions. *Differential Diagnosis of Autism Spectrum Disorder* (2022), 1.

[74] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–18. doi:10.1145/3411764.3445432

[75] 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. doi:10.1145/3290605.3300831

[76] Liuping Wang, Zhan Zhang, Dakuo Wang, Weidan Cao, Xiaomu Zhou, Ping Zhang, Jianxing Liu, Xiangmin Fan, and Feng Tian. 2023. Human-centered design and evaluation of AI-empowered clinical decision support systems: a systematic review. *Frontiers in Computer Science* 5 (June 2023), 1187299. doi:10.3389/fcomp.2023.1187299

[77] Geoffrey Webb, Claude Sammut, Claudia Perlich, Tamás Horváth, Stefan Wrobel, Kevin Korb, William Noble, Christina Leslie, Michail Lagoudakis, Novi Quadrianto, Wray Buntine, Lise Getoor, Galileo Namata, Jiawei Jin, Jo-Anne Ting, Sethu Vijayakumar, Stefan Schaal, and Luc De Raedt. 2010. Leave-One-Out Cross-Validation.

[78] Mary K. Wojczynski and Hemant K. Tiwari. 2008. Definition of Phenotype. In *Advances in Genetics*. Vol. 60. Elsevier, 75–105. doi:10.1016/S0065-2660(07)00404-X

[79] Sara Wolf, Tobias Grundgeiger, Raphael Zähringer, Lora Shishkova, Franzisca Maas, Christina Dilling, and Oliver Happel. 2025. How a Clinical Decision Support System Changed the Diagnosis Process: Insights from an Experimental Mixed-Method Study in a Full-Scale Anesthesiology Simulation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–23. doi:10.1145/3706598.3713372

[80] Siyi Wu, Weidan Cao, Shihan Fu, Bingsheng Yao, Ziqi Yang, Changchang Yin, Varun Mishra, Daniel Addison, Ping Zhang, and Dakuo Wang. 2025. CardioAI: A Multimodal AI-based System to Support Symptom Monitoring and Risk Prediction of Cancer Treatment-Induced Cardiotoxicity. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–22. doi:10.1145/3706598.3714272

[81] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–11. doi:10.1145/3290605.3300468

[82] Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M. Padilla, Jeffrey Caterino, Ping Zhang, and Dakuo Wang. 2024. Rethinking Human-AI Collaboration in Complex Medical Decision Making: A Case Study in Sepsis Diagnosis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–18. doi:10.1145/3613904.3642343

[83] Yujie Zhu, Mingyuan Zhang, Cong Fang, Le Fang, Meichen Liu, Yonghao Long, Kun-Pyo Lee, Lie Zhang, and Stephen Jia Wang. 2025. AI Doctor for ASD: Physician Perceptions and Adoption Challenges in Autism Clinical Practice. *Proc. ACM Hum.-Comput. Interact.* 9, 2 (May 2025), CSCW027:1–CSCW027:28. doi:10.1145/3710925

A Formative Study Interview Protocol

Below we provide an outline of the interview phases.

- **Phase 1**: Questions about their professional experience, current work focus, describing their diagnosis process, and current challenges
- **Phase 2**: Introducing the SIT, Questions about prior knowledge
- **Phase 3**: Open questions about non-verbal behavior of interest in the SIT context
- **Phase 4**: Exploration and interpretation of schematic visualization of the SIT video data
- **Phase 5**: Question about the possible integration of the SIT into practice
- **Phase 6**: Questions about demographics and concluding information

B User Study Interview Protocol

Below we provide an outline of the interview phases. Phase 3 was done twice, first with the Non-ASC case, and then with the ASC case.

- **Phase 1**: Questions about their professional experience with a focus on ASC
- **Phase 2**: Introduction to the SIT-CARE
- **Phase 3.1**: Video-based assessment of case
  - Initial assessment of non-verbal behavior with confidence statement
  - Describe own reasoning
  - Predict model-based assessment and reasoning
  - Predict what the data-based assessments could show for specific non-verbal behavior
- **Phase 3.2**: Data-based assessment of case
  - Reevaluate diagnosis with confidence statement
  - Describe own reasoning
  - Agreement with data-based assessments
  - Predict model-based assessment and reasoning
- **Phase 3.3**: Model-based assessment of case
  - Reevaluate diagnosis with confidence statement
  - Describe own reasoning
  - Assume reasoning of model
- **Phase 4**: Questions about comprehension, predictability, usefulness and trustworthiness of outputs
- **Phase 5**: Questions about what they liked, disliked and would like to improve
- **Phase 6**: Questions about integration into workflow
- **Phase 7**: Attitudes towards AI

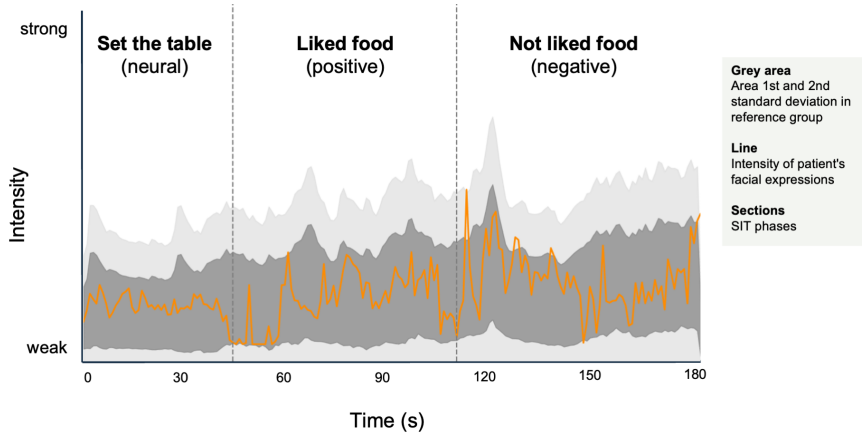

Fig. 3. Visualization of the chosen facial expression, intensity and variability of positive facial expressions per phase, of a schematic Non-ASC case.

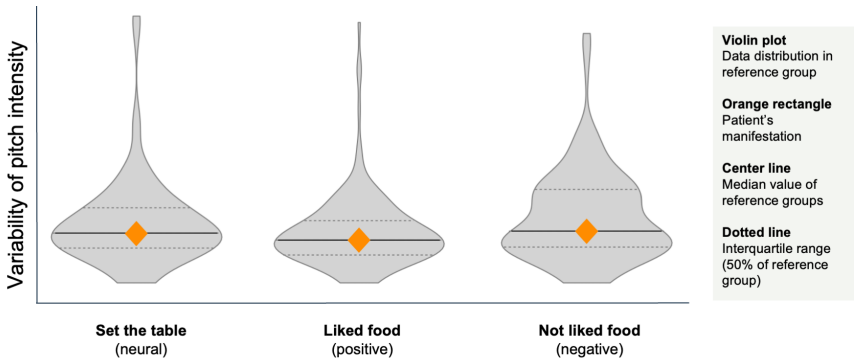

Fig. 4. Visualization of the chose voice parameter, variability of pitch, of a schematic Non-ASC case.

C Data-based visualization

D Detailed Calculations for Data-based Analysis

This appendix provides the detailed computational procedures for Section 4.2.

D.1 Gaze behavior

Fixation density levels. To create the heat maps of gaze fixations, raw gaze vectors were projected from the webcam coordinate system to a screen-centered coordinate space. Fixation density was color-coded into five discrete levels, defined relative to a reference group (RG, Non-ASC participants) as:

$$\text{Intervals}_{\text{RG}} = \begin{cases} (-\infty,\ \mu_{\text{RG}} - 1.5\sigma_{\text{RG}}) & \text{very low fixation density} \\ [\mu_{\text{RG}} - 1.5\sigma_{\text{RG}},\ \mu_{\text{RG}} - 0.5\sigma_{\text{RG}}) & \\ [\mu_{\text{RG}} - 0.5\sigma_{\text{RG}},\ \mu_{\text{RG}} + 0.5\sigma_{\text{RG}}) & \\ [\mu_{\text{RG}} + 0.5\sigma_{\text{RG}},\ \mu_{\text{RG}} + 1.5\sigma_{\text{RG}}) & \\ (\mu_{\text{RG}} + 1.5\sigma_{\text{RG}},\ \infty) & \text{very high fixation density} \end{cases}$$

where μ_{RG} and σ_{RG} denote the mean and standard deviation of fixation counts in the reference group.

Representative examples. To illustrate prototypical gaze distributions, we pooled all frame-level gaze points for ASC and Non-ASC groups separately. From each pool, we randomly sampled 5,500 frames (approx. equal to the number of frames per participant) using a fixed random seed (`random_state=42`) to ensure reproducibility. These sampled frames were then visualized as heat maps.

Numerical statistics. In addition to visualizations, the following statistics were computed for each assessed participant:

(1) Variance of horizontal gaze angles.
(2) Percentage of time spent looking at the screen.
(3) Percentage of time directed toward the actress's face region.

For comparison, mean and standard deviation of these metrics were computed for both ASC and Non-ASC reference groups.

D.2 Facial expressions

Representative examples. Representative individuals were identified using a distance-to-median score:

$$\text{score}_i = \frac{\text{total\_abs\_diff}_i}{\text{median}_j(\text{total\_abs\_diff}_j)} + \sum_{p \in \{\text{neutral, joy, disgust}\}} \frac{\left|\text{Var}_i^{(p)} - \text{median}_j\left(\text{Var}_j^{(p)}\right)\right|}{\text{median}_j\left(\left|\text{Var}_j^{(p)} - \text{median}_k\left(\text{Var}_k^{(p)}\right)\right|\right)}.$$

Here, total_abs_diff_i is the sum of absolute deviations between subject i's trace and the group median trace, while $\text{Var}_i^{(p)}$ denotes variance within phase p. The two subjects with the lowest scores were visualized as prototypical examples. For anonymization, 15% Gaussian noise was added to the traces.

Numerical statistics. For each participant, we computed:

(1) Mean and variance of smiling intensity (AU06 + AU12).
(2) Mean and variance of AU04 (brow lowering, associated with negative affect).
(3) Overall mean intensity across all AUs.

Reference distributions (mean ± SD) were calculated separately for ASC and Non-ASC groups, and additionally for gender-specific subgroups.

D.3 Voice parameters

For voice features, we extracted from the openSMILE eGeMAPS set:

(1) Variance of pitch (`F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2`) (representing the range of pitch (in semitones, relative to 27.5 Hz) between the 2nd and 98th percentiles of voiced segments, thereby minimizing the effect of outliers).

(2) Mean and variance of voiced segment length (`MeanVoicedSegmentLengthSec`, `StddevVoicedSegmentLengthSec`^{2}).
(3) Loudness variability (`loudness_sma3_stddevNorm`^{2}).

Prototypical examples corresponded to the median values of ASC and Non-ASC groups. Group-level means and standard deviations were computed for ASC and Non-ASC reference groups. Visualizations of pitch variability were provided via violin plots, stratified by interaction phase, while numerical results were displayed alongside the assessed participant's values.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009