How Can I Signal You To Trust Me: Investigating AI Trust Signalling in Clinical Self-Assessments

Naja Kathrine Kollerup
nkka@cs.aau.dk
Aalborg University
Aalborg, Denmark

Joel Wester
joelw@cs.aau.dk
Aalborg University
Aalborg, Denmark

Mikael B. Skov
dubois@cs.aau.dk
Aalborg University
Aalborg, Denmark

Niels van Berkel
nielsvanberkel@cs.aau.dk
Aalborg University
Aalborg, Denmark

ABSTRACT

Individuals are increasingly interested in and responsible for assessing their own health. This study evaluates a fictional AI dermatologist for assistance in the self-assessment of moles. Building on the Signalling Theory, we tested the effect of textual descriptions provided by a virtual dermatologist, as manipulated across 'Ability', 'Integrity, ' and 'Benevolence', along with the clinical assessment, 'benign' or 'malignant', affect users' trust in the aforementioned trust pillars. Our study (N = 40) follows a 2 (Ability low/high) × 2 (Integrity low/high) × 2 (Benevolence low/high) × 2 (mole assessment benign/malignant) within-subject factorial design. Our results demonstrate that we can successfully influence perceptions of ability and benevolence by manipulating the corresponding aspects of trust but not perceived integrity. Further, in the case of a malignant assessment, participants' perception of trust increased across all aspects. Our results provide insights into the design of AI support systems for sensitive use cases, such as clinical self-assessments.

KEYWORDS

Human-AI Trust Building, Trust, Signalling theory, AI Self-Assessment Tools

ACM Reference Format:
Naja Kathrine Kollerup, Joel Wester, Mikael B. Skov, and Niels van Berkel. 2024. How Can I Signal You To Trust Me: Investigating AI Trust Signalling in Clinical Self-Assessments. In *Designing Interactive Systems Conference (DIS '24), July 01–05, 2024, IT University of Copenhagen, Denmark.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3643834.3661612

1 INTRODUCTION

The integration and development of artificial intelligence (AI) systems have taken a major role in assisting humans in a variety of areas of social and health-related aspects [53], such as clinical decision-making [14, 64, 73] and patient self-care [7]. Integrating AI-support systems as a collaborator for medical professionals is often seen as a solution to reduce pressure on the healthcare system and overworked personnel [50, 58]. This extensive pressure on our healthcare system has also resulted in a gradual increase of patient taking responsibility for their own clinical self-care, such as the monitoring of glucose levels by diabetics [51] or the use of self-care tools for people with Parkinson's disease [39]. The use of AI-support systems by patients in their self-care practices results in a shift in clinical AI collaboration, forcing non-experts to engage with AI systems. This raises numerous questions regarding the use of AI systems to support patients in assessing their health—including *how* these systems can be designed to elicit trust.

Acknowledging the importance of trust, research efforts are now directed at designing trustworthy AI-support systems to assist patients in their clinical self-care [1]. While the integration of AI in healthcare shows promise in increasing both efficiency and quality of clinical care [45, 51], it also reveals challenges in understanding the shifting dynamics of trust and responsibility from healthcare professionals to patients. From this, patients can be required to rely on their understanding and interpretation of AI-provided information, transitioning away from face-to-face interactions with medical professionals towards computer-mediated communication [67]. This shift in communication distances the observable factors in our behaviour–mimics, facial expressions, and body language. These observable factors provide us with signals that convey information about the behaviours and intentions of others, which is crucial in establishing and maintaining trust. This opens up an interesting space for designers to investigate how AI-support systems can convey trust signals, allowing end-users to increase or decrease their trust where appropriate.

In this paper, we draw on the Signalling Theory, which describes how signals are sent to convey information to others [61]. Specifically, we investigate how the trust pillars of ability, benevolence, and integrity (ABI), traditionally centred in human-to-human trust building [47], can be manipulated as trust-enhancing signals in AI systems designed for clinical self-assessment. We investigate the impact of the absence or presence of ABI in a realistic clinical self-care scenario: self-assessing moles for skin cancer. Specifically, we seek to answer the following research question: *How do varying levels of trust signals from an AI system impact users' self-reported trust-building behaviours and interactions with an AI-support system?* To answer this question, we follow a 2 (Ability low/high) ×

DIS '24, July 01–05, 2024, IT University of Copenhagen, Denmark

ACM ISBN 979-8-4007-0583-0/24/07
https://doi.org/10.1145/3643834.3661612

2 (Benevolence low/high) × 2 (Integrity low/high) × 2 (Mole assessment Benign / Malignant) within-subject factorial design (N = 40). We designed a mock AI dermatologist, which presents a set of assessments and recommendations of moles to our participants. We evaluated the effect on participants' compliance with the recommendation and self-reported trust measures.

We find that manipulating the three trust pillars effectively alters participants' perceptions of 'ability' and 'benevolence' but not 'integrity'. The presentation of a malignant (cancerous) assessment as compared to a benign assessment notably increased participants' trust across all trust pillars. Our manipulation of the three trust pillars—Ability, Benevolence, and Integrity (ABI)—did not influence participants' compliance with the AI recommendations. However, we observed that images clinically assessed as malignant did affect participants' willingness to follow the AI's recommendations.

Through this work, we provide insights for designing AI trust signals—specifically in support systems for sensitive use cases, such as clinical self-assessments. We emphasise that designers should focus on *adjusting provided AI assistance to high and low-stake scenarios*, given that the relevance of ABI as trust-building pillars can differ across settings, influencing users' trust-building. Further, *account for patients' medical history*, since prior health experiences can significantly impact individuals' trust. Lastly, AI-assisted self-assessment tools should *act as supportive rather than authoritative assistance.*

2 RELATED WORK

We motivate our work through prior work highlighting trust as key to human-human interactions and a prerequisite for successful human-AI interaction (e.g. [2, 38, 66]). Furthermore, we take inspiration from the Signalling Theory to investigate trust in human-AI interactions.

2.1 Trust Factors

Trust is a multifaceted concept—making it challenging to integrate into AI systems to improve patient experiences. A common inspiration for designing AI systems to increase trust can be taken from trust in human-human interactions. Mayer et al. defines interpersonal trust as *"the willingness of A to be vulnerable to the actions of B based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party"* [47, p. 712].

A significant amount of research within human-AI trust-building relies on Mayer's proposed definition of trust [47, 66]—with further examples such as behavioural perceptions of trust [28, 31] and ABI as trust measurement questionnaire [22]. Furthermore, ABI has recently been used to investigate trust in human-AI interactions. For example, Hauptman et al. recently explored the use of ABI in designing and implementing trustworthy AI systems [63]. Similarly, Centeio Jorge et al. explore how an AI system can assess human teammates as trustworthy by observing human behaviour through ABI in an online experiment [35], and Jakesh et al. who used ABI to investigate the trustworthiness of profile text written by an AI [31].

Within this established notion of trust in human-human and human-AI interactions— conceptualising, designing, and studying the effects of ABI signals on people's trust towards AI systems is a promising way forward.

2.2 Trust Signalling

When going from face-to-face communication to human-AI communication, the aspect of nonverbal cues (*e.g.* facial expressions, body language, and eye contact) is reduced, which allows for a more open interpretation of information where direct observation is not possible [67]. These unobservable factors relate to questions around trust *e.g. when to trust the other individual—when observable factors are not possible, but forced to construct knowledge from unobservable factors.* [4]. Bacharach et al. suggest three concepts to describe these factors: *manifesta, i.e. observable features*, such as facial expressions and behaviour, and *krypta, i.e. unobservable features*, such as ABI [4]. In trusting others, these concepts are relied upon when forming opinions about the qualities of the other individual—mediating knowledge of non-directly observable qualities [4, 23]. In the context of human-AI interactions, Jorge et al. highlight the need for a more profound understanding of internal characteristics (e.g., ability, benevolence, and integrity), and how these can be discerned through observable behaviours [16]. Falcone et al. suggest that these observable behaviours act as signals of the underlying qualities, which are able to explain the behaviours in specific interactions [24].

To further an understanding of the role of these signals, Signalling Theory can shed further light on how *(krypta)* the unobservable trust pillars (ABI) [47] can be signalled in a human-AI context—and if these signals can be observed in the behaviour of the AI *(manifesta)*. The Signalling Theory describes and categorises the communication between two parties–signaller and receiver. The *Signaller* signals information to the *Receiver*, which interprets the signal, and feedback is returned to the *Signaller* [20, 61, 62]. The focal point in this information transaction is to convey the *Receiver* to act accordingly to the signals being sent and provide actions that work in the favour of the *Signaller*.

The Signalling Theory has previously been investigated in HCI to understand the interaction between people as mediated by technology. Lampe et al. used the Signalling Theory to investigate how social media profiles signal specific elements for articulating user connections and relationships. Their findings indicate that popular profiles are associated with the number of friends signalled in the profile fields [41]. Furthermore, Warner et al. recently used the Signalling Theory framework to understand how HIV statuses are being disclosed in dating applications [69]. Their findings suggest that participants preferred to keep their status undisclosed, developing signalling appropriation strategies. In contrast, other participants developed counter-signals (e.g., reducing the unravelling effect) to minimise privacy [69]. Shami et al. build on the Signalling Theory to investigate people's interpretation of information in online profiles in evaluating expertise [57]. They highlight that specific signals (e.g., social connection info), particularly in online profiles—were considered more reliable indicators of expertise according to their participant's decision of whom to contact [57].

We understand the Signalling Theory as a conceptual framework, aligning with the theoretical perspective of Spence [61], and similar research conducted by Shami et al. [57]. Building on this framework,

we investigate the dynamics of trust-building between humans and AI within the context of AI-assisted health self-assessment tools. Further, exploring how nonverbal cues (i.e. ability, benevolence, and integrity) can be conveyed as signals through observed textual behaviour to increase human-AI trust.

In this study, an AI system (the Signaller) would emit signals that reflect its ability, integrity, and benevolence. The users (the Receivers) interpret these signals, influencing their trust in the AI-support system. The core of these interactions lies in the unobservable (krypta)–ABI–of the AI to be signalled through observable behaviours (manifesta), thus guiding the users to respond in ways that are beneficial for both parties involved. The challenge and focal point here is ensuring that the signals sent by the AI are accurately received and interpreted by users, leading to trust-based actions.

2.3 Trust and Human-AI Interactions

Building on recent research on trust in human-AI interactions [28, 31, 44], we are particularly interested in these interactions in a healthcare context—including clinical-decision making [14, 73], patient diagnosis [15], and prognosis [9] in critical and vulnerable settings. When introducing AI-support systems to healthcare settings, we observe challenges in AI systems taking on the role of a collaborator, assisting medical personnel in decision-making [14, 73]. However, a shift is observed when AI takes the role of a medical expert consulting and assisting patients in their treatment. Thus, it is unclear how AI-support systems can be designed to increase trust so patients are comfortable with being assisted in the scenarios of AI-assisted health self-assessment tools.

Berge et al. explore how to design AI systems supporting nurses in their clinical assessment through collaboration with nurses through co-design workshops and interview [10]. The authors urge designers to consider design for document support (standardisation of complex and varied problems) and document automation (allowing for flexibility). Furthermore, Kaltenhauser et al. investigate design opportunities for clinical decision support systems in intensive care through a field study exploring how physicians and nurses collaborate for optimal care [36]. The authors discuss design considerations centred around enhancing user interface adaptability, improving collaboration in decision-making, maintaining transparency, and upholding both human-centred interaction and data integrity in machine learning applications [36].

Recent work has explored human-AI trust-building mechanisms. For example, Buçina et al. focused on reducing over-reliance in AI-assisted decision-making by comparing three cognitive forcing designs: withholding AI suggestions, requiring initial user decisions without AI assistance, and delaying AI recommendations [13]. Their findings indicate that cognitive forcing effectively diminishes over-reliance on AI systems compared to immediate exposure to AI suggestions.

Similarly, Panigutti et al. examined how AI explanations can foster trust, particularly in healthcare. They compared a baseline AI-suggestion condition against an enhanced one that included patient history explanations. Despite participants expressing dissatisfaction with the explanation quality, their reliance on AI advice significantly increased when explanations were provided [54]. Their research showed that without model performance data, individuals' reliance on an AI model is significantly influenced by their agreement with the model on high-confidence decisions. However, this effect changes when aggregated performance information of the model is provided, and the level of individual confidence moderates this reliance, especially in cases of disagreement with the model's recommendations [54]. These prior works rely on the ability or performance of AI systems when investigating user's trust. However, we suggest broadening the scope of AI's ability to fully cover the multifaceted trust concept. In this context, we suggest that *benevolence* [28] and *integrity* [49], as inspired by human-human trust literature [47], is a promising way forward regarding trust-building in human-AI interactions.

Since late 2022, researchers and developers have been exploring how LLMs can be used and how users currently use them for various tasks. These tasks range from straightforward question-answering to more open-ended conversations without a clear purpose [59]. For example, Purohit et al. showed how ChatGPT can support individuals with language disorders in retrieving words they have on the tip of the tongue [55]. While LLMs generally show impressive capabilities for such tasks, it is unclear how LLMs are used and what tasks are particularly relevant to users. Recently, Wang et al. investigated how conversational agents can be powered by LLMs for conversational interactions [68], mainly focusing on how LLMs can power mobile UIs. While understanding how LLMs can power mobile applications for different purposes remains underexplored, Wang et al. present a promising way forward in how to use LLMs in the design of (e.g. health-focused) applications. One clear benefit of LLMs is the natural language format, making LLMs accessible to non-experts. However, using natural language to steer the behaviour of LLMs is not an easy task, which Zamfirescu et al. recently showed by investigating how non-experts go about in prompt engineering [74]. Others have explored how different strategies for prompts can positively impact LLM behaviours. Tongshuang et al. recently introduced LLM chaining, where the LLM output guides the shape of the following LLM output, ultimately increasing controllability for end-users in shaping LLM output [71].

While the papers as mentioned above focus on how users can access LLMs, often limited to developers and other technical experts, others focus on assessing the applicability of LLMs in various contexts (e.g., health and well-being). Jo et al. recently developed an LLM-powered chatbot designed to support socially isolated individuals, where their results point to both promises and challenges for using LLMs to support such individuals (and other LLM-powered interactive applications) through empathic, text-based conversations [33]. This has been further emphasised by Xiao et al., focusing on how experts can inform the design of credible health information to information seekers during COVID-19. Together with others, they developed the AI chatbot Jennifer and evaluated it with both experts and information seeking, providing a foundation for how AI systems can be 'expert-sourced' to work as a stable alternative for accessing health information during crisis [72]. While LLMs show promise in a range of domains and tasks, challenges and risks are increasingly highlighted in HCI [30, 42, 75]. In this paper, informed by ABI and Signalling Theory, we aim to evaluate the effects of LLM-generated textual presentations of clinical self-assessments.

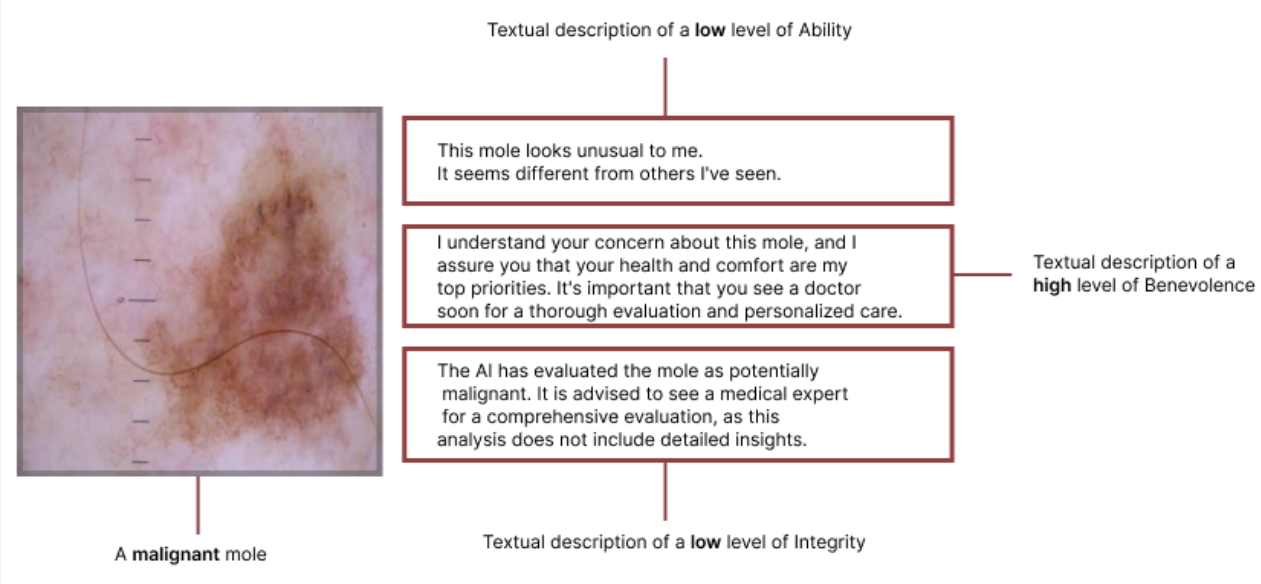

Figure 1: Task as shown to a participant with additional annotations in red. The AI's assessment presented with a low level of ability, high level of benevolence, and low level of integrity.

3 METHOD

We designed an online study that mimics a self-assessment scenario to assess the impact of different trust signals on user trust behaviour. Specifically, we seek to test the following three hypotheses related to participant perception of an AI support tool.

H1: When exposed to signals emphasising a high ability, high integrity, and high benevolence, participants will exhibit higher levels of trust in an AI-based medical assessment as compared to exposure to low ability, low integrity, and low benevolence signals.

H1 follows from Mayer's observation that a high level of the three trust pillars leads to a trustee deemed as trustworthy [47]. As stated, trust should be considered a continuum, where the three pillars can vary along this continuum and consequently be affected by context [47]. Following this, we hypothesise that decision-making in sensitive or vulnerable situations can affect users' trust perception. From this, we introduce the following hypotheses:

H2: The positive impact of a high benevolence signal will be higher when presented with a malignant assessment as compared to a benign assessment.

H2 follows from the literature on sensitive patient care, which states that emotional care with patients in sensitive situations is critical to comfort patients, e.g., when faced with chronic illnesses [40]. This supports H2, given that high benevolence is an integrated part of communication to support patients in vulnerable situations.

H3: The positive impact of a high ability signal will be higher when presented with a benign assessment as compared to a malignant assessment.

If the AI's assessment suggests that a mole is benign and does not require further follow-up with a professional, participants may feel less pressure to fully accept the result without question. Highlighting the AI's strong diagnostic abilities reassures the user that a highly capable system has properly evaluated even innocuous-looking moles. This increased reassurance boosts trust in the benign assessment more so than for a malignant assessment, where participants likely feel a greater need for additional reassurance or professional follow-up.

3.1 Experimental design

Our study follows a 2 (Ability low/high) × 2 (Integrity low/high) × 2 (Benevolence low/high) × 2 (Mole assessment Benign / Malignant) within-subject factorial design. For this study, we designed a mock AI dermatologist that assesses moles and provides a medical recommendation (e.g., consult with a human doctor or not) through textual descriptions.

In real-world clinical settings, dermatologist examine the moles for abnormalities, such as changes in colour, size, shape or texture. Should any irregularities be detected, a dermatologist may perform a biopsy, removing a sample of tissue for microscopic examination to determine the presence of cancer cells [29]. A considerable minority of atypical moles may develop into melanoma. Moreover, while the lifetime risk of melanoma for individuals is less than 1%, this risk escalates to over 10% for those with atypical moles [37].

3.1.1 Manipulation of trust signals. We sought to create a reproducible yet realistic AI-generated conversation to assess the impact of manipulating trust signals in a clinical assessment. Specifically, we manipulated the levels of ABI to represent either a low or high definition, as summarised in Table 1. We adopt Mayer's definition of ability as "*the group of skills, competencies, and characteristics that enable the trustee to influence the domain*", and benevolence as "*the extent to which a trustee wants to do good to the trustor, aside from*

Level	Ability	Integrity	Benevolence
Low	Limited medical knowledge and experience.	Suggests a lack of transparency and honesty regarding its medical recommendations.	Appears indifferent or uncaring about the patient's well-being, potentially implying a self-serving agenda.
High	Emphasises its extensive medical knowledge. Assess the images and point out areas that look suspicious, portraying itself as a highly competent and experienced virtual doctor.	Emphasises its commitment to honesty, transparency, and awareness of risks and biases.	Expresses genuine concern for the patient's well-being, promising to prioritise their health and comfort.

Table 1: An overview of the low and high definitions of ability, benevolence, and integrity. We prompted ChatGPT-4 with these definitions and requested a low and high description of each.

an egocentric profit motive" [47]. Lastly, we adopt the definition of *integrity* from Mehrotra et al., who define integrity as "*honesty, transparency and fairness in sharing risks such as biases*" [49] – a definition often used in the human-AI literature [49].

Our study features a total of 16 tasks, with eight benign and eight malignant moles randomly distributed between these tasks. The number of tasks was determined by the total number of possible combinations of low and high levels of *ABI*.

For each task, participants were shown an image of a mole and a text snippet describing the mole as benign or malignant as per the combination of low/high trust signals. The text snippet in each task was consistently presented in the following order: image assessment (representing the ability signal), the recommendation provided (representing the benevolence signal), and assessment clarification (representing integrity signal), as seen in Figure 1. This presentation order was informed by the RESPECT model (rapport, empathy, support, partnership, explanations, cultural competence, and trust) used in doctor-patient communication, emphasising respect and openness towards patients [3]. As such, the task presentation mirrored real-world doctor-patient interactions. The images of the benign and malignant moles were selected from a widely used melanoma dataset [32], of which ground truth information is available. We ensured that the presented AI assessment always aligned with the ground truth assessment of the mole.

After being presented with the image and generated text-based recommendation, participants were asked to either accept or decline the recommendation provided by the AI dermatologist as to following or not following up with a doctor consultation, an established approach to studying trust in Human-AI collaboration [18, 65].

To ensure the validity of our medical recommendations and images, we engaged with a dermatology expert to ensure an accurate assessment of the showcased mole images in the study. We presented the dermatologist with the selected images and AI-generated recommendations. Figure 2 shows an example of a final prompt and the generated output, describing a low ability and a clinical assessment of a benign mole.

Prompt instruction: Participants will be shown images of a malignant or benign mole. The text shown to the participants will present an assessment of the image. The first text should represent an assessment of a benign mole. It should be written from a perspective of low ability. Low ability is defined as the following: Limited medical knowledge and experience. The parameters that need to be considered in the text: Extremely generic descriptions of the benign mole image, the text must be two sentences, not reflect on own personal skills, do not mention any of the moles characteristics.

Output: This mole looks pretty normal to me. It appears typical and not out of the ordinary.

Figure 2: Example of GPT-4 prompt instruction (assessment of a benign mole, written from a low ability perspective) and the generated output.

3.1.2 GPT-4 textual manipulations. The text snippets, as presented to participants, were generated using OpenAI's ChatGPT-4 (gpt-4-0125-preview, temperature of 1.0) to ensure a realistic AI interaction. Since each task featured three short text snippets (Figure 1), each of which could represent a high or low level of their respective trust factor, across a total of eight tasks, we required 48 unique text snippets to avoid the chance of exact repetition between tasks for the same participant. Examples of text snippets from malignant assessments generated by GPT-4 are shown in Table 2, while those from benign assessments are shown in Table 3.

Using GPT-4, we generated unique text snippets, matching the levels (low or high) of the trust pillars and the clinical assessment (benign or malignant).

We first prompted GPT-4 with the definitions of low or high levels of ABI, as described in Table 1, combined with a benign or malignant clinical assessment. Each prompt round was explicitly focused on a specific trust pillar (ability, benevolence or integrity), the definition level (low or high), and the clinical assessment (benign or malignant).

We iteratively refined the prompts for GPT-4, adjusting and removing elements from our prompt until the text successfully covered the critical aspects of ABI, see Table 1. Following the iteration

Trust signals	High	Low
Ability	This image displays a mole with asymmetry, uneven borders, and variegated colors, characteristics commonly associated with malignancy. The irregular diameter and evolving size suggest potential malignant transformation	This mole looks unusual compared to others. It seems different in color and size
Benevolence	I can see why this mole might be worrying you, and your health and comfort are my utmost concern. Please make an appointment to see a doctor, so we can take the best care of you	It seems necessary for you to schedule a visit with a doctor regarding this mole
Integrity	The AI's evaluation suggests malignancy in the mole, identified through patterns commonly seen in cancerous lesions. It is important to consider the limitations of AI, including potential data biases and the necessity of a professional medical diagnosis for accurate assessment	This AI-generated analysis indicates a significant concern with the mole. Please note, further details on the medical recommendation are not available at this time

Table 2: An overview of samples illustrating low and high representations of ability, benevolence, and integrity for a malignant image.

Trust signals	High	Low
Ability	Observing the mole's uniform coloration and symmetrical borders, it aligns with typical benign characteristics. Its smooth, regular shape and absence of irregularities suggest a non-cancerous nature, typically meaning the mole is harmless and not indicative of skin cancer	From what I can see, this looks like a normal skin spot. It doesn't seem unusual to me
Benevolence	I understand your concerns and want to assure you that your health is my primary focus. At present, a visit to the doctor isn't necessary	Seeing a doctor for this seems quite unnecessary
Integrity	This analysis, conducted by an AI system, indicates that the mole is benign. Please note that while AI provides a high degree of accuracy, it is not infallible and should be supplemented with professional medical evaluation	Our AI system has processed the image and indicates the mole is benign. For detailed medical advice, please consult a healthcare professional

Table 3: An overview of samples illustrating low and high representations of ability, benevolence, and integrity for a benign image.

rounds, we constructed the tasks by extracting text snippets from ABI, adding them to the respective structure—RESPECT model—and assigning the correct levels, low or high, needed for the unique tasks.

3.2 Measurements

To assess participants' perceived trust, we used trusting belief items adapted from McKnight et al. [48], initially developed by Mayer [46]. This questionnaire, incorporating Mayer's definitions of trust, is well-established in the literature on human-AI trust-building [66]. It consists of 11 questions divided between the three areas of trust (ABI), with answers on a 5-point Likert scale ranging from 'Strongly disagree' to 'Strongly agree'. Participants answered these questions for each of the 18 tasks.

To assess participants' attitudes towards the system, we incorporated an open-ended question following prior work [66]. Participants were asked: "*Imagine getting a consultation from a virtual doctor. What do you consider as important factors in a virtual doctor (e.g., showing capabilities, showing honesty and transparency, showing consideration of your well-being)? Please elaborate on your answer in the text box below:*". Participants answered the questions at the end of completing the survey.

Finally, we collected demographic information (age, gender) and whether participants or their loved ones had dealt with skin cancer – to assess the impact of participants' prior experience on the topic, potentially explaining variations in trust [66].

3.3 Participants

We conducted an *a priori* power analysis using G*Power version 3.1 [25]. Using an effect size of 0.2, categorised as small [19], a significance level of $\alpha = .05$, and a desired power of 0.95, the calculated minimum sample size was $N = 36$. Our participants were recruited through Prolific ($N = 40$), with criteria including an approval rate above 95% and being native English speakers. We recruited exclusively Caucasian participants, following careful consideration and internal debate. A critical aspect of our study involves visually examining moles on the skin. Due to limitations in the availability of diverse dermatological image data sets [21], the images we were able to source and use predominantly featured moles on Caucasian skin types. As the visual characteristics of moles can vary significantly across different skin types, we chose to solely recruit participants with a skin tone similar to the images used. We did not restrict participants to specific geographical locations. We ensured that participants could only participate once.

Variable	Estimate	SE	Statistic	*p*-value	
AI manipulations					
Ability Level (High)	-0.83	0.71	-1.17	0.240	
Benevolence Level (High)	-0.26	0.74	-0.35	0.723	
Integrity Level (High)	0.04	0.73	0.05	0.958	
Clinical Assessment (Malignant)	3.12	0.48	6.46	< 0.001	***
Participant characteristics					
Skin cancer (Yes)	0.35	0.66	0.53	0.594	
Interaction effects					
Ability Level (High):Benevolence Level (High)	1.95	1.06	1.85	0.064	
Ability Level (High):Integrity Level (High)	-0.05	0.99	-0.05	0.961	
Benevolence Level (High):Integrity Level (High)	-1.09	1.01	-1.08	0.280	
Ability Level (High):Benevolence Level (High):Integrity Level (High)	-0.87	1.42	-0.61	0.540	

Table 4: An overview of the predictive variables and interaction effects for participants recommendation adherence.

4 RESULTS

A total of 40 participants (15 males, 23 females, and 1 non-binary) participated in the experiment. We excluded one participant who failed our attention check. The mean age of these participants was 39 years old, ranging between 21 and 68 years of age. Participant age distribution was as follows: 21-30 (N = 10), 31-40 (N = 13), 41-50 (N = 8), 51-60 (N = 5), and 61-68 (N = 3) years. Participants were recruited from a wide range of countries, including Australia (2), Ireland (4), Spain (11), Canada 3, Israel (1), New Zealand (2), United Kingdom (19), Greece (1), Mexico (1), South Africa (1) and the USA (3). The average completion time was approximately 15 minutes.

4.1 Model construction

To test our three hypotheses, we constructed three generalised mixed-effect models (GLMM) using the R package *lme4* [8]. Initially, we constructed the model by including eight predictive variables (Ability level, Benevolence level, Integrity Level, Skin Cancer, Virtual Consultation, Gender, Age, and Clinical Assessment). Following, we removed variables based on their Akaike information criterion (AIC) in a step-wise approach. Our final predictor selection resulted in a total of five predictive variables for our models. We specified Response ID as a random effect in all predictive models to account for individual differences between participants. We investigated the impact of the following predictive variables.

- **Clinical assessment:** describes the clinical assessment presented to the participants. Images were either benign or malignant pictures of moles.
- **Ability, Benevolence, and Integrity level:** describes the defined level–low or high–of the generated text-based instructions. See Table 1.
- **Skin cancer:** characterises participants' personal experience with skin cancer, binary variable.

4.2 Recommendation adherence

In 227 out of 312 possible recommendations, participants decided to follow the recommendations, follow– or not to follow up with a doctor consultation provided by the AI system, corresponding to 72%. A Chi-square test shows no statistically significant evidence to suggest a difference in adherence rates to the AI recommendation between the participants' gender categories (χ^2 = 55, df = 2, p = 0.759). Next, we present the Decision model, in which we assess the effect of the aforementioned predictors on participants' compliance with the provided recommendations.

We conducted a likelihood test between our final model and the null model to test the goodness of fit [11]. The results from the likelihood test show that our model is statistically significant (χ^2 = 76, p < 0.001), accounting for 62% of the variance in the response variable (R^2 = 0.628). Finally, we tested for multicollinearity among the models' variables and found a variation inflation factor (VIF) between 1.00 and 7.57 for our predictors. From this, we can indicate that the values are below the generally used threshold of ten for detecting multicollinearity [27].

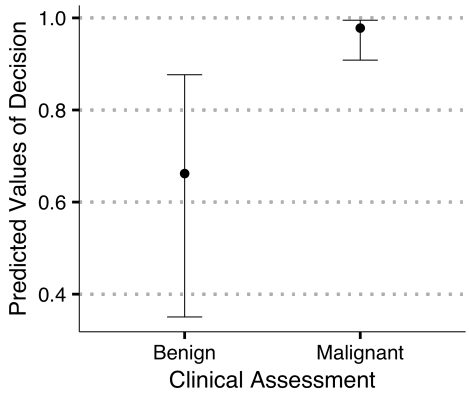

Figure 3: Main effect of clinical assessment on participants' decision to follow the recommendation provided.

We find that 'Clinical Assessment' significantly impacts the outcome variable 'Decision'. We do not find a significant effect of the other predictors. Table 4 provides an overview of these results. The interaction between 'Ability level' and 'Benevolence level' is close to significant p = 0.06. Figure 3 shows the impact of clinical

Variable	Estimate	SE	Statistic	*p*-value	
AI manipulations					
Ability Level (High)	0.28	0.15	1.92	0.002	**
Benevolence Level (High)	0.18	0.15	1.22	0.87	
Integrity Level (High)	0.35	0.15	2.40	0.59	**
Clinical Assessment (Malignant)	0.41	0.08	5.15	<0.001	***
Participant characteristics					
Skin cancer (Yes)	0.40	0.19	2.17	0.036	*
Interaction effects					
Ability Level (High):Benevolence Level (High)	0.03	0.21	0.15	0.85	
Ability Level (High):Integrity Level (High)	-0.37	0.21	-1.79	0.304	
Benevolence Level (High):Integrity Level (High)	-0.64	0.21	-3.05	0.005	**
Ability Level (High):Benevolence Level (High):Integrity Level (High)	0.44	0.29	1.51	0.132	

Table 5: An overview of the predictive variables and interaction effects for participants' perceived ability.

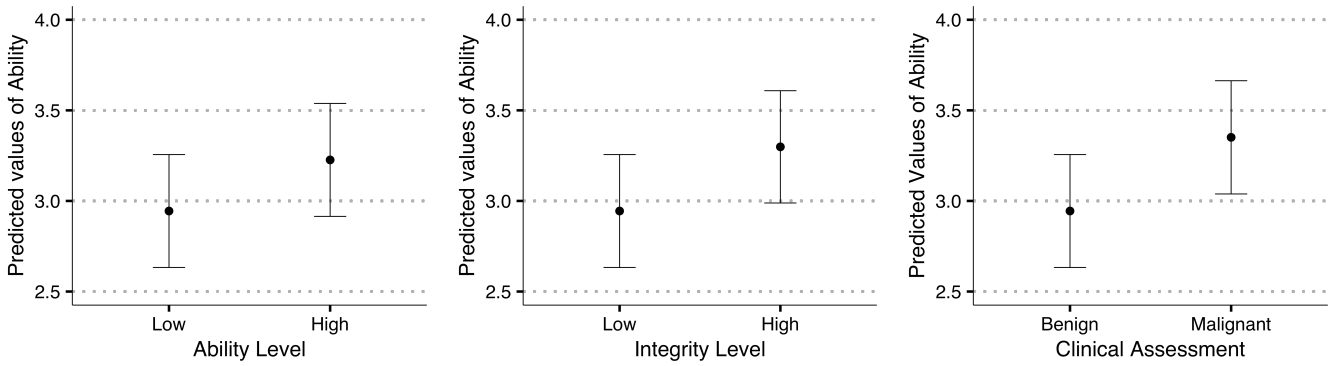

Figure 4: Main effects of ability (left), integrity (centre), and clinical assessment (right) on participants' perceived ability.

assessment on participants' compliance with the AI recommendation. We observe that when participants were shown a task of a 'Malignant' assessment, participants were more likely to follow the recommendation provided by the AI system. Participants followed the recommendations 139 times out of 227 when shown malignant images, equivalent to 61.2%. In comparison, for the benign assessments, participants followed the recommendations 38.8% of times – a difference of 22.4%pt.

4.3 Ability perceptions

First, we present the Ability model to assess the effect of the aforementioned predictive variables on participants' perceived ability of the AI dermatologist. To test the goodness of fit of our predictive model, we conducted a likelihood ratio test between our final model and the null model. These results show that our model is statistically significant ($\chi^2 = 50$, $p < 0.001$), accounting for 47.5% of the variance in the response variable ($R^2 = 0.475$). We tested for the existence of multicollinearity and found a VIF between 1.00 and 7.00 for our predictors – below the commonly used threshold of ten for detecting multicollinearity [27].

Table 5 shows that the main effects of ability level, integrity level, clinical assessment, and participant's history with skin cancer are significant predictors of perceived ability. In addition, we find a significant interaction effect between benevolence level and integrity level. Figure 4-left shows the predicted values of perceived ability based on a low and high AI ability level – with a low AI ability resulting in a significantly lower perceived ability. A similar effect is found for the low and high integrity levels, Figure 4-centre. Furthermore, we find a main effect of clinical assessment, as visualised in Figure 4-right, with malignant assessments receiving a higher perceived ability as compared to benign images.

Figure 5-left shows the main effect of participants' having personally dealt with skin cancer and their self-reported ability trust score. Finally, we find a two-way interaction effect between benevolence level and integrity level, as shown in Figure 5-right. First, the plot highlights that when the integrity level is low, participants' perceived ability is higher when the benevolence level is high. With a high level of integrity, the perceived ability is higher when the benevolence level is low. This indicates that the relationship between benevolence levels and integrity levels inverts to a higher level of integrity.

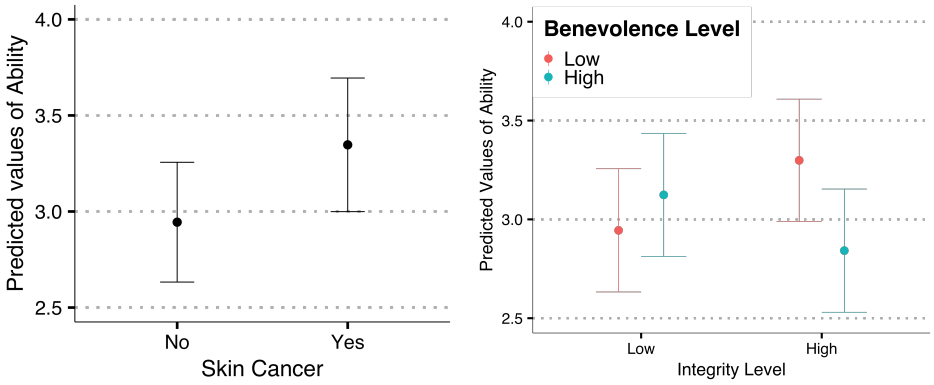

Figure 5: Main effects of participants' history with skin cancer (left) and interaction effect between benevolence and integrity level (right) on participants' perceived ability.

4.4 Benevolence perceptions

We present the Benevolence model to assess the effect of the aforementioned predictive variables on participants' perceived benevolence of the AI dermatologist. To test the goodness of fit of our predictive model, we conducted a likelihood ratio test between our full model and the null model [11]. These results show that our model is statistically significant ($X^2 = 24.7, p < 0.003$) with a 62% variance in the response variable ($R^2 = 0.620$). We tested the existence of multicollinearity and found a VIF between 1.00 and 7.00 for our predictors – below the commonly used threshold of ten for detecting multicollinearity [27].

Table 6 show that the main effects of clinical assessment and benevolence level are significant predictors of perceived benevolence. Figure 6 shows the predicted values of perceived benevolence based on a low and high AI benevolence level – with a low AI benevolence resulting in a significantly lower perceived benevolence. In addition, we find a main effect of clinical assessment, as visualised in Figure 7, with malignant assessment receiving a higher perceived benevolence as compared to benign images.

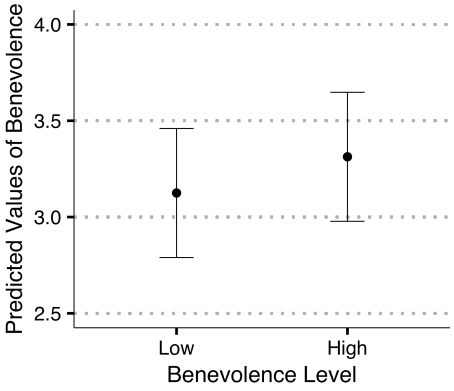

Figure 6: Main effects of benevolence on participants' perceived benevolence.

Variable	Estimate	SE	Statistic	*p*-value	
AI manipulations					
Ability Level (High)	-0.03	0.13	-0.20	0.989	
Benevolence Level (High)	0.19	0.13	1.49	0.010	*
Integrity Level (High)	0.17	0.13	1.31	0.336	
Clinical Assessment (Malignant)	0.23	0.07	3.39	< 0.001	***
Participant characteristics					
Skin cancer (Yes)	0.40	0.23	1.78	0.083	
Interaction effects					
Ability Level (High):Benevolence Level (High)	0.15	0.18	0.82	0.626	
Ability Level (High):Integrity Level (High)	-0.01	0.18	-0.07	0.448	
Benevolence Level (High):Integrity Level (High)	-0.11	0.18	-0.64	0.116	
Ability Level (High):Benevolence Level (High):Integrity Level (High)	-0.17	0.25	-0.67	0.506	

Table 6: An overview of the Predictive variables and Interactions effects in the Benevolence model.

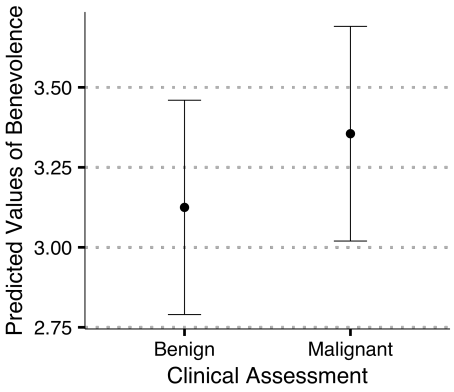

Figure 7: Main effects of clinical assessment on participants' perceived benevolence.

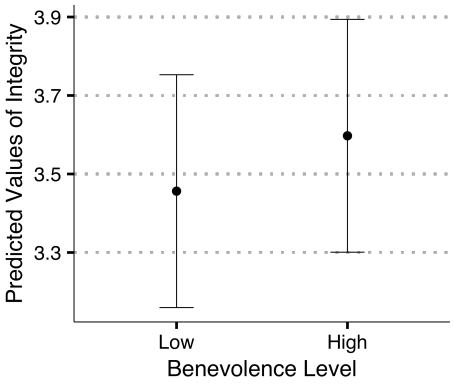

Figure 8: Main effects of benevolence on participants' perceived integrity.

4.5 Integrity perceptions

We present the Integrity model to assess the effect of the aforementioned predictive variables on participants' perceived integrity of the AI dermatologist. To test the goodness of fit for our predictive model, we conducted a likelihood ratio test between our full model and the null model [11]. These results show that our model is statistically significant (χ^2 = 18.3, p = 0.028) with a 58.3% variance in the response variable (R^2 = 0.583)). We tested the existence of multicollinearity and found a VIF between 1.00 and 7.00 for our predictors–below the commonly used threshold of ten for detecting multicollinearity [27].

Table 7 show that the main effects of clinical assessment and benevolence level are significant predictors of perceived integrity. Figure 8 shows the predicted values of integrity based on low and high AI benevolence levels – with a high AI benevolence resulting in a significantly higher perceived integrity. Furthermore, we find a main effect of clinical assessment as visualised in Figure 9, with a malignant assessment receiving higher perceived integrity as compared to a benign assessment.

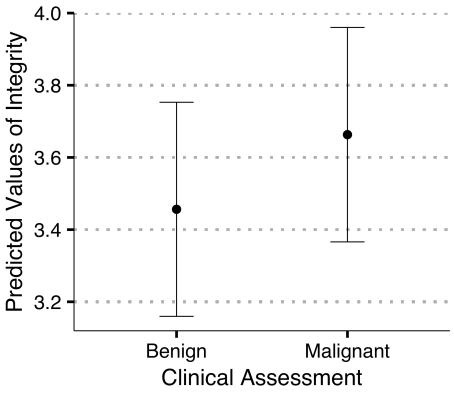

Figure 9: Main effects of clinical assessment on participants' perceived integrity.

Variable	Estimate	SE	Statistic	*p*-value	
AI manipulations					
Ability Level (High)	0.12	0.11	1.07	0.204	
Benevolence Level (High)	0.14	0.11	1.23	0.026	*
Integrity Level (High)	0.09	0.11	0.77	0.641	
Clinical Assessment (Malignant)	0.21	0.06	3.35	< 0.001	***
Participant characteristics					
Skin cancer (Yes)	0.10	0.20	0.48	0.632	
Interaction effects					
Ability Level (High):Benevolence Level (High)	0.00	0.16	0.00	0.969	
Ability Level (High):Integrity Level (High)	-0.09	0.16	-0.58	0.391	
Benevolence Level (High):Integrity Level (High)	-0.02	0.16	-0.14	0.808	
Ability Level (High):Benevolence Level (High):Integrity Level (High)	-0.01	0.23	-0.04	0.969	

Table 7: An overview of the Predictive variables and Interactions effects in the Integrity model.

4.6 Qualitative analysis

After completing every task, participants answered an open-ended question about what factors in the AI-based recommendation affected their view towards the AI. We sought insights into how the text-generated recommendations containing low or high signals of ABI impacted participants' trust. In addition, at the survey's conclusion, participants answered an open-ended question about what they consider important factors in a virtual doctor.

We systematically coded participants' responses following a deductive thematic analysis approach [12]. From this, we aimed to understand the participants' perception of ABI, when incorporated in the recommendation as signals in the AI system. When conducting the deductive thematic analysis, we developed a coding framework based on themes derived from our research objectives: the impact of ABI on AI trust signalling. Firstly, we familiarised ourselves with the data and identified themes, and then afterwards, we labelled segments of participants' responses to the respective themes.

4.6.1 ***The impact of Ability on AI trust signalling.*** Through our deductive thematic results, we identified three codes involving the AI support systems Ability; Detailed explanations, Clinical detection, and Knowledgeable. Our results highlight that '*Detailed explanations*' are deemed as highly important by our participants:"*This diagnosis was detailed enough and well written, I would take the suggestion to see a doctor*" (P1). These results are emphasised by the contradicting responses from the participants, highlighting that lack of detail made them question the AI; "*The AI seemed unsure and did not give a detailed analysis. This made me question the competency of the AI*" (P10). Similarly, the participants highlighted the AI as '*knowledgeable*'; "*I would be reassured and happy with the AI explanation as they knew what they were talking about*" (P39). In addition, participants emphasised the AI system's ability to '*Clinically detect*' the characteristics of benign or malignancy in the moles as positive: "*It listed several different examples of why it believed the mole to be non-cancerous, these could be identified and confirmed in the image*" (P28).

4.6.2 ***The impact of Benevolence on AI trust signalling.*** Three codes were identified within Benevolence: Empathy, Lack of empathy, and Consulting with a doctor. Our findings highlight '*Empathy*' as a focal point for the participants when asked what affected their view towards the AI. Within this, the participants used words such as *Concerning*, *Sincere*, and *caring* when describing the behaviour of the AI; "*The answer showed empathy and gave me a sense of comfort like it wanted to help me and showed me compassion*" (P14). Another stating: "*Is sincere and recommends the best for me*" (P4). These results are emphasised by the contradicting responses from the participants, highlighting that '*Lack of empathy*' worried the participants in relation to receiving bad news: "*It would be quite worrying to receive this potential news from an AI that shows no empathy at a particularly worrying time*" (P14). In addition, the majority of the participants had positive reactions when the AI recommended them to '*Consult with a doctor*', as a follow-up check; "*This message was very well written, it had empathy but also again suggested getting further assessment if you think it's needed*" (P34).

4.6.3 ***The impact of Integrity on AI trust signalling.*** Two codes were identified within Integrity: Honesty and Fairness. Our findings indicate that participants describe the AI as honest, and the '*Honesty*' around the limitations in the AI was perceived positively by the participants; "*The advice was good, and noted that it should not be relied upon. I would be happy with this AI advice*" (P20). However, we also observe contradicting statements towards this, with participants stating that the AI being fallible made them question its trustworthiness; "*The statement about AI being fallible made me question the accuracy of the recommendation*" (P27). In addition, a participant described the AI as *Fair* concerning what affected their view; "*I am not a fan of AI but felt this time comments were fair and accurate to consider further*" (P4).

4.6.4 ***Important factors in a virtual doctor.*** Within categorising factors the participants consider important in a virtual doctor, we identified three codes: Empathy, Knowledgeable, and Honesty. The majority of the statements from the participants are positioned within the themes of '*Empathy*' and '*Knowledgeable*'. Our findings highlight that the participants deemed empathy in a virtual doctor as highly important, focusing specifically on acknowledgement of concerns, compassionate, and considerate of their well-being; "*They listen and understand your concerns - they take their time to explain what is happening - they are concerned for your welfare*" (P5). Furthermore, knowledge was stressed as an important element, especially the ability to explain the diagnosis; "*I think being able to explain why they have come up with a diagnosis, rather than just saying it is what it is*" (P29). Concerning '*Honesty*', participants highlighted a virtual doctor needs to be honest with the patient; "*Honest is vital as your health is the most important thing. They also need to have a caring manner, or else they are working in the wrong profession*" (P39). Honesty as a character trait is often described in combination with empathy and knowledge, where the participants are stressing the importance of transparency and empathy; "*Transparency definitely - not pretending to be an expert that can replace doctors and to always assure people that it's fine if they also want to see a doctor. Not to make it sound like it they do they are wasting people's time. Showing empathy and understanding in the language is important*" (P26).

5 DISCUSSION

Our results offer insights into trust building in human-AI interaction, particularly in designing for sensitive use cases such as clinical self-assessments. We find that both high levels of ability and benevolence increase participants' self-reported trust scores. Furthermore, a malignant assessment influences the participants' self-reported trust scores across all three trust pillars: ABI. Further, a prior history of skin cancer influenced participants' perception of the AI dermatologist's ability. Finally, when presented with a malignant assessment, participants are more likely to follow the AI's recommendation than when presented with a benign assessment.

5.1 Effects of Context on Perceptions of AI Trust Signals

Our results show that images clinically assessed as malignant increased the self-reported trust score within each ABI pillar. Further, participants reported a higher trust score in the AI dermatologist's ability if they have had a history of skin cancer (personally or close

family), as seen in Figure 5 (left). These results confirm **H2** concerning benevolence; "*The positive impact of a high benevolence signal will be higher when presented with a malignant assessment as compared to a benign assessment.*." We did not hypothesise that images assessed as malignant would also increase participants' trust in benevolence and integrity, thereby rejecting **H3**; "*The positive impact of a high ability signal will be higher when presented with a benign assessment as compared to a malignant assessment*".

These findings indicate that the severity of the clinical assessment significantly impacted participants' self-reported trust, also enhancing their adherence to the recommendations provided. This highlights how a specific context (e.g., prior experiences and how concerning news is communicated) can significantly influence end-users' trust towards an AI system. Our results align with the framework 'Health Belief Model', which suggests that people's health-related behaviours are influenced by their perceptions of the severity of a health threat [34]. This framework suggests that patients are increasingly more likely to take action to prevent illness if they believe it would have severe consequences for their health—a phenomenon known as perceived severity [34].

Building on this, it is critical to examine how health information is communicated in clinical settings, especially when delivering bad news. Inappropriate communication methods can negatively influence patients' perceptions of their illness [60]. Doctors' lack of gentleness, carefulness, and time in interactions leads to negative emotions among patients [60]. Prior work by Kwon et al. highlights that in sensitive care, empathy and listening to patients' needs is critical [40]. Our results emphasises this, as benevolence increased the participants' self-reported trust measures. Further, the participants highlighted that empathy positively affected their view towards the AI, as stated in Section 4.6.2. Furthermore, in high-stakes situations, such as treatment decisions for malignant moles, the perception of risk significantly influences our decision-making process. Trust does not emerge without a sense of vulnerability, such as having something significant at stake, which plays a crucial role in influencing decision-making processes [66]. Our findings similarly show that participants' trust in the AI system's ability and benevolence increases when exposed to malignant images. This highlights how vulnerability impacts trust in AI-assisted systems when contrasted with lower-stake health scenarios. We discuss the ethical considerations of (trust in) AI-assisted self-assessment in Section 5.3.

5.2 Influence of ABI pillars on Trust Development in AI-Healthcare Interactions

We found that the presence of high trust signals ability and benevolence positively impacted participants' self-reported trust levels. These results demonstrate that manipulating the corresponding aspects of trust can successfully impact the participants' perceptions of ability and benevolence. These results partially confirm **H1**; "*When exposed to signals emphasising a high ability, high integrity, and high benevolence, participants will exhibit higher levels of trust in an AI-based medical assessment as compared to exposure to low ability, low integrity, and low benevolence signals.*" We did not observe this effect for integrity and the respective trust score. We did not observe any significant effects between the ABI pillars and the participants' compliance in following the recommendations.

Prior work shows that perceived ABI can change over time [47, 49]. This raises the question of the effect of ABI on trust building in longer interactions. Mayer et al. state that benevolence is a specific attachment and an emotional connection that will increase over time as the relationship between the trustor and trustee develops [47]. This suggests that an increase in benevolence might be difficult to realise in short-term human-AI interactions. Our findings contradict this, as we observe that benevolence significantly affects participants' self-reported trust scores. This discrepancy might result from our study's focus on a relatively high-stake scenario of clinical self-assessment. Previous work by Kwon et al. [40] highlights that in sensitive care, emphasising high benevolence (e.g., showing care and listening to the patient's needs) is deemed highly important for terminally ill patients. As such, a high level of benevolence and ability might be deemed critical for designing AI-support systems in clinical settings.

While high levels of ABI are synonymous with users' trust-building, the essential objective of any human-AI interaction should be to appropriately calibrate user trust and avoid over- or under-trust. Prior work highlights the importance of a calibrated approach to trust building [13, 52]. We, similarly, do not argue that ABI should always be high. As stated by Mayer et al., ABI is affected by the context, and the perception of ABI will change as the context of the task being performed is changed [47]. For example, a perceived high ability at one task does not necessarily imply a high perceived ability at another task. Mayer et al. emphasise the challenge of how low the trust signals can be before one is deemed as untrustworthy [47].

5.3 Ethical Considerations of AI-assisted Health Self-assessment Tools

AI systems for clinical self-assessments can help assist individuals in their everyday lives. However, it is critical to consider the associated risks that arise in high-stakes contexts. This is particularly important when AI is used as a self-assessment tool to aid patients with their healthcare outside of traditional clinical environments—arguably a high-stakes scenario. According to Nunes et al., patients make decisions related to their care every day; however, these decisions are not made within a uniform context, given that risks can vary significantly [51]. Therefore, as the authors argue, it is crucial to reflect on the importance of discerning which decisions should involve clinician support and which can be managed more autonomously by patients [51].

Based on our results, we further emphasise that the risk of AI-assisted health self-assessment tools relates to the distinction of autonomy levels. A majority of our participants valued the AI recommending them to consult with a doctor. From this, participants had a positive assessment towards the AI, with participants deeming the follow-up consultation with a doctor important. Baldauf et al. studied perceptions of AI-driven self-diagnosis apps and found that while users were positive towards the concept in close combination with general practitioner care, they had reservations about this as a stand-alone concept [5]. Therefore, both from an ethical and user perspective, it is critical to design AI systems to act in accordance with clinical experts.

As a consequence of AI being increasingly integrated into society, algorithmic harm (i.e., AI decisions with negative or harmful outcomes) also receives increased attention [6, 17, 43]. When reflecting on the potential algorithmic harm in medical self-assessment *outside* of clinical settings, a focal point revolving around the risks of assessments is AI systems' accuracy, both in terms of output quality and how that output quality is communicated. When designing trustworthy AI-support systems, a critical aspect is under-trust in clinical AI support. Additionally, there is a significant risk that individuals may place too much trust in AI, potentially postponing seeking professional medical advice when necessary. Within the context of AI-assisted health self-assessment tools, however, the impact of false positives may be deemed less critical than the consequences of not adopting these as the latter can lead to missed opportunities for early detection and intervention.

5.4 Design Considerations for Human-AI Trust Signalling

Non-embodied AI systems, such as text or image-based clinical self-assessment tools, cannot make use of non-verbal cues in their interactions with humans. This lack of nonverbal cues results in an open interpretation of communication by the recipient [67]. Consequently, we argue that alternative signals can be used as indicators of non-directly observable qualities [23]. Through trust signalling, we can communicate attributes such as ability, benevolence, and integrity. These attributes are communicated through signals that make these qualities observable in the behaviour of the AI systems [4]. Similar to Jorge et al. [35], investigating the ABI as observable features in the behaviour of humans in collaboration with AI agents lead to an effect on the overall trustworthiness in all three ABI pillars. Based on our results, we suggest three concrete design considerations related to manipulating user trust.

DC1: AI Self-assessment Tools Should Adjust Provided Support For High-stake Scenarios. We recommend that designers carefully assess the specific contexts, given that sensitive use cases, such as malignancy, affected participants' compliance and trust in following the recommendations provided by the AI system. This is explained as 'perceived severity', underlining that if a person believes their condition to be severe, they will take action to prevent illness [34]—indicating that users may be more motivated to 'blindly follow' AI recommendations in high-stake scenarios. Consequently, designers must fine-tune how information, particularly bad news, is communicated, as individuals are more likely to trust and adhere to the guidance provided. Given the potential for over-trust, careful consideration must be given to how these tools foster trust without undermining the need for professional medical consultation. To tackle this issue, we encourage designers to manipulate integrity in AI systems to openly acknowledge the AI's limitations and the accuracy of its assessments.

Additionally, designers should consider the broad spectrum of AI applications, from high-stakes environments such as health self-assessments to everyday, low-risk situations. It is essential to acknowledge that these contexts shape end-users' trust in AI-assisted systems, as discussed in Section 5.1. Research within doctor-patient communication states that empathy, attention, sharing information, and understanding patient perspectives are critical concepts within doctor-to-patient communication [26]. Building on these principles, we argue for AI-assisted health self-assessment tools in high-stake scenarios to communicate the system's ability by using straightforward medical terminology, providing detailed explanations, and highlighting abnormalities. Benevolence can be expressed through empathetic language, considerate responses, and recommendations to consult a doctor. Our qualitative research aligns with studies in doctor-patient communication, which suggest that respect, honesty, and attentiveness enhance patients' trust in their doctors [70].

DC2: AI Self-assessment Tools Should Account For Patients' Medical History. When designing AI-assisted health self-assessment tools, it is crucial for designers to take into account the prior medical history of participants. Our findings highlight that a patient's previous health experiences (i.e., skin cancer) can significantly impact their trust in the system's ability, see Figure 5. Mayer's work on trust dynamics highlights how ABI as trust-building pillars can differ across settings, influencing users' trust-building [47]—adapting the communication of ABI levels could be necessary for the AI to adapt to specific situations.

Based on these insights, we recommend that designers prioritise communicating a high level of ability, see Section 5.4, particularly when designing for patients with an elaborate medical history. While benevolence and integrity remain important, they should be calibrated appropriately to ensure that the primary focus on ability does not get overshadowed by these other crucial elements of trust. Additionally, we advise designers to evaluate participants' previous medical experiences before commencing the study. This aligns with previous research [66], which indicates that prior experiences can impact trust in AI-supported systems, as demonstrated in our findings, see Figure 5.

DC3: AI-assisted Self-assessment Tools Should Act As a Supportive Rather Than Authoritative Assistance. To enhance trust in AI-assisted self-assessment tools during the initial screening phase, the design of these tools should clearly acknowledge that AI is not a substitute for medical consultation. With this comes the considerations of the level of control the AI have within the treatment decision in self-assessment tools. Our findings indicate that participants valued recommendations from the AI to consult with a doctor, and such guidance significantly increased their trust in the AI system—aligning with prior research on the utilisation of AI in clinical practices [56]. Additionally, considering the influence of authority bias, we recommend that designers position AI-assisted self-assessment tools explicitly as supportive rather than authoritative assistance. It should be clearly communicated prior to the interaction that the AI is not a substitute for professional medical advice, as highlighted in Section 4.6.4.

5.5 Limitations and Future Work

We acknowledge several limitations in our work. First, while we investigated imagined self-assessment scenarios, participants were unaffected by the assessments. As such, our results could differ from those of assessments that directly impacted participants' lives. Second, different scenarios may require higher or lower trust by participants, for example, when the stakes of the decision are higher or lower. Therefore, we cannot generalise our findings to all scenarios. Furthermore, we limited participant recruitment to Caucasian

individuals. This decision was driven by the specific requirements of our experimental design, particularly the nature of the visual materials used. The images of skin conditions, central to our research, were not aligned with specific needs for the study [32]. Due to this limitation in image diversity, we chose to recruit participants whose skin type matched the images to ensure consistency and accuracy in the assessment and interpretation of these conditions. Finally, future work may consider how human-AI trust building develops over time, explore different low- or high-stakes scenarios that might affect the pillars of ABI, and explore how these can be manipulated to accommodate appropriate trust.

6 CONCLUSION

This paper investigates the impact of ability, integrity, and benevolence on participants' trust towards an AI-support system. We explored the impact of low and high levels of these three trust pillars in a realistic clinical self-assessment scenario: assessing moles. We follow a 2 (Ability low/high) × 2 (Integrity low/high) × 2 (Benevolence low/high) × 2 (Mole assessment Benign / Malignant) within-subject factorial design (N = 39). we designed a mock AI dermatologist to assess pictures of benign or malignant moles, and based on the assessment, the AI dermatologist provides recommendations to the participant. We assessed the effect on participants' compliance with the recommendation and self-reported trust measures. Our findings showed that manipulating different aspects of trust can successfully influence participants' perceptions of ability and benevolence. Further, clinical assessment of malignant images increased the participants' self-reported trust score in all three trust pillars. Surprisingly, we observed an increase in the ability trust score if the participants had a prior history of skin cancer. Through our results and design recommendations, we provide insights into the design of AI support systems for sensitive use cases, such as clinical self-assessments.

ACKNOWLEDGMENTS

This work is supported by Digital Research Centre Denmark (DIREC) project 'Explain-Me' under Innovation Fund Denmark.

REFERENCES

[1] Onur Asan, Zhongyuan Yu, and Bradley H Crotty. 2021. How clinician-patient communication affects trust in health information sources: Temporal trends from a national cross-sectional survey. *PLoS One* 16, 2 (2021), e0247583.

[2] Maryam Ashoori and Justin D. Weisz. 2019. In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes. *CoRR* abs/1912.02675 (2019). https://doi.org/abs/1912.02675

[3] Dawn M Aycock, Traci T Sims, Terri Florman, Karis T Casseus, Paula M Gordon, and Regena G Spratling. 2017. Language sensitivity, the RESPECT model, and continuing education. *J. Contin. Educ. Nurs.* 48, 11 (2017), 517–524. https://doi.org/10.3928/00220124-20171017-10

[4] Michael Bacharach and Diego Gambetta. 2001. *Trust in Signs.* 148–184. https://psycnet.apa.org/record/2001-16661-005

[5] Matthias Baldauf, Peter Fröehlich, and Rainer Endl. 2020. Trust Me, I'm a Doctor – User Perceptions of AI-Driven Apps for Mobile Health Diagnosis. In *Proceedings of the 19th International Conference on Mobile and Ubiquitous Multimedia (MUM '20).* Association for Computing Machinery, New York, NY, USA, 167–178. https://doi.org/10.1145/3428361.3428362

[6] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 74 (apr 2021), 34 pages. https://doi.org/10.1145/3449148

[7] Matthew Barrett, Josiane Boyne, Julia Brandts, Hans-Peter Brunner-La Rocca, Lieven De Maesschalck, Kurt De Wit, Lana Dixon, Casper Eurlings, Donna Fitzsimons, Olga Golubnitschaja, Arjan Hageman, Frank Heemskerk, André Hintzen, Thomas M Helms, Loreena Hill, Thom Hoedemakers, Nikolaus Marx, Kenneth McDonald, Marc Mertens, Dirk Müller-Wieland, Alexander Palant, Jens Piesk, Andrew Pomazanskyi, Jan Ramaekers, Peter Ruff, Katharina Schütt, Yash Shekhawat, Chantal F Ski, David R Thompson, Andrew Tsirkin, Kay van der Mierden, Chris Watson, and Bettina Zippel-Schultz. 2019. Artificial intelligence supported patient self-care in chronic heart failure: a paradigm shift from reactive to predictive, preventive and personalised care. *EPMA J.* 10, 4 (2019), 445–464. https://doi.org/10.1007/s13167-019-00188-9

[8] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[9] Maura Bellio, Dominic Furniss, Neil P. Oxtoby, Sara Garbarino, Nicholas C. Firth, Annemie Ribbens, Daniel C. Alexander, and Ann Blandford. 2021. Opportunities and Barriers for Adoption of a Decision-Support Tool for Alzheimer's Disease. *ACM Trans. Comput. Healthcare* 2, 4, Article 32 (2021), 19 pages. https://doi.org/10.1145/3462764

[10] Arngeir Berge, Frode Guribye, Siri-Linn Schmidt Fotland, Gro Fonnes, Ingrid H. Johansen, and Christoph Trattner. 2023. Designing for Control in Nurse-AI Collaboration During Emergency Medical Calls. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS '23).* Association for Computing Machinery, New York, NY, USA, 1339–1352. https://doi.org/10.1145/3563657.3596110

[11] Benjamin M. Bolker, Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* 24, 3 (2009), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

[12] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* 57–71. https://doi.org/10.1037/13620-004

[13] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (2021), 21 pages. https://doi.org/10.1145/3449287

[14] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. *Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making.* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300234

[15] 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. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (2019), 24 pages. https://doi.org/10.1145/3359206

[16] Carolina Centeio Jorge, Myrthe L. Tielman, and Catholijn M. Jonker. 2022. Artificial Trust as a Tool in Human-AI Teams. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI '22).* IEEE Press, 1155–1157. https://doi.org/10.1109/HRI53351.2022.9889652

[17] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23).* Association for Computing Machinery, New York, NY, USA, 651–666. https://doi.org/10.1145/3593013.3594033

[18] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 370 (2023), 32 pages. https://doi.org/10.1145/3610219

[19] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences* (2 ed.). Routledge, London, England. https://doi.org/10.4324/9780203771587

[20] Brian L Connelly, S Trevis Certo, R Duane Ireland, and Christopher R Reutzel. 2011. Signaling theory: A review and assessment. *J. Manage.* 37, 1 (2011), 39–67. https://doi.org/10.1177/0149206310388419

[21] Roxana Daneshjou, Kailas Vodrahalli, Roberto A. Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M. Swetter, Elizabeth E. Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, Johan A. C. Allerup, Utako Okata-Karigane, James Zou, and Albert S. Chiou. 2022. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances* 8, 32 (2022), eabq6147. https://doi.org/10.1126/sciadv.abq6147 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.abq6147

[22] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems That Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI Explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23).* Association for Computing Machinery, New York, NY, USA, Article 352, 13 pages. https://doi.org/10.1145/3544548.3580672

[23] Judith Donath. 2007. *Signals, Truth & Design.* MIT Press, Chapter Signals, cues and meaning (draft), 1–34.

[24] Rino Falcone, Michele Piunti, Matteo Venanzi, and Cristiano Castelfranchi. 2013. From Manifesta to Krypta: The Relevance of Categories for Trusting Others. *ACM Trans. Intell. Syst. Technol.* 4, 2, Article 27 (2013), 24 pages. https://doi.org/10.1145/2438653.2438662

[25] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* 39, 2 (2007), 175–191. https://doi.org/10.3758/bf03193146

[26] Jennifer Fong Ha and Nancy Longnecker. 2010. Doctor-patient communication: a review. *Ochsner J.* 10, 1 (2010), 38–43.

[27] Joseph F Hair, William C Black, Barry J Babin, Rolph E Anderson, and RL Tatham. 2010. *Multivariate Data Analysis.* Pearson.

[28] Allyson I. Hauptman, Wen Duan, and Nathan J. Mcneese. 2022. The Components of Trust for Collaborating With AI Colleagues. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing (CSCW'22 Companion).* Association for Computing Machinery, New York, NY, USA, 72–75. https://doi.org/10.1145/3500868.3559450

[29] Dana-Faber Cancer Institute. 2024. *How We Diagnose Melanoma, Homepage.* https://www.dana-farber.org/cancer-care/types/melanoma/diagnosis#:~:text=A%20skin%20exam%20checks%20for,to%20check%20for%20cancer%20cells Online; accessed 3-May-2024.

[30] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23).* Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. https://doi.org/10.1145/3544548.3581196

[31] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. 2019. AI-Mediated Communication: How the Perception That Profile Text Was Written by AI Affects Trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300469

[32] Muhammad Hasnain Javid. 2022. Melanoma Skin Cancer Dataset of 10000 Images. https://doi.org/10.34740/KAGGLE/DSV/3376422

[33] Eunkyung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23).* Association for Computing Machinery, New York, NY, USA, Article 18, 16 pages. https://doi.org/10.1145/3544548.3581503

[34] Christina L Jones, Jakob D Jensen, Courtney L Scherr, Natasha R Brown, Katheryn Christy, and Jeremy Weaver. 2015. The Health Belief Model as an explanatory framework in communication research: exploring parallel, serial, and moderated mediation. *Health Commun.* 30, 6 (2015), 566–576.

[35] Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. 2023. How Should an AI Trust Its Human Teammates? Exploring Possible Cues of Artificial Trust. *ACM Trans. Interact. Intell. Syst.* (2023). https://doi.org/10.1145/3635475 Just Accepted.

[36] Annika Kaltenhauser, Verena Rheinstädter, Andreas Butz, and Dieter P. Wallach. 2020. "You Have to Piece the Puzzle Together": Implications for Designing Decision Support in Intensive Care. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference (DIS '20).* Association for Computing Machinery, New York, NY, USA, 1509–1522. https://doi.org/10.1145/3357236.3395436

[37] Patrick M. Zito Katherine E. Wensley. 2023. *Atypical Mole.* https://www.ncbi.nlm.nih.gov/books/NBK560606/#:~:text=A%20significant%20minority%20of%20atypical,moles%2C%20greater%20than%2010%25. Online; accessed 3-May-2024.

[38] Bran Knowles, Mark Rouncefield, Mike Harding, Nigel Davies, Lynne Blair, James Hannon, John Walden, and Ding Wang. 2015. Models and Patterns of Trust. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15).* Association for Computing Machinery, New York, NY, USA, 328–338. https://doi.org/10.1145/2675133.2675154

[39] Elina Kuosmanen, Eetu Huusko, Niels van Berkel, Francisco Nunes, Julio Vega, Jorge Goncalves, Mohamed Khamis, Augusto Esteves, Denzil Ferreira, and Simo Hosio. 2023. Exploring crowdsourced self-care techniques: A study on Parkinson's disease. *International Journal of Human-Computer Studies* 177 (2023), 103062. https://doi.org/10.1016/j.ijhcs.2023.103062

[40] Sinyoung Kwon, Miyoung Kim, and Sujin Choi. 2020. Nurses' experiences of providing "sensitive nursing care" for terminally-ill individuals with cancer: A qualitative study. *European Journal of Oncology Nursing* 46 (2020), 101773. https://doi.org/10.1016/j.ejon.2020.101773

[41] Cliff A.C. Lampe, Nicole Ellison, and Charles Steinfield. 2007. A Familiar Face(Book): Profile Elements as Signals in an Online Social Network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07).* Association for Computing Machinery, New York, NY, USA, 435–444. https://doi.org/10.1145/1240624.1240695

[42] Q. Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. (2023). https://doi.org/10.48550/arXiv.2306.01941

[43] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2023. Blaming Humans and Machines: What Shapes People's Reactions to Algorithmic Harm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23).* Association for Computing Machinery, New York, NY, USA, Article 372, 26 pages. https://doi.org/10.1145/3544548.3580953

[44] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23).* Association for Computing Machinery, New York, NY, USA, Article 759, 19 pages. https://doi.org/10.1145/3544548.3581058

[45] Gergely Magyar, João Balsa, Ana Paula Cláudio, Maria Beatriz Carmo, Pedro Neves, Pedro Alves, Isa Brito Félix, Nuno Pimenta, and Mara Pereira Guerreiro. 2019. Anthropomorphic Virtual Assistant to Support Self-care of Type 2 Diabetes in Older People: A Perspective on the Role of Artificial Intelligence. In *VISIGRAPP.* https://api.semanticscholar.org/CorpusID:88480572

[46] C. Roger Mayer and H. James Davis. 1999. The Effect of the Performance Appraisal System on Trust for Management: A Field Quasi-Experiment. (1999).

[47] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (1995), 709–734. http://www.jstor.org/stable/258792

[48] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Inf. Syst. Res.* 13, 3 (2002), 334–359. https://doi.org/10.1287/isre.13.3.334.81

[49] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. 2023. Integrity Based Explanations for Fostering Appropriate Trust in AI Agents. *ACM Trans. Interact. Intell. Syst.* (2023). https://doi.org/10.1145/3610578 Just Accepted.

[50] Dana L Nelson-Peterson and Carol J Leppa. 2007. Creating an environment for caring using lean principles of the Virginia Mason Production System. *J. Nurs. Adm.* 37, 6 (2007), 287–294. https://doi.org/10.1097/01.NNA.0000277717.34134.a9

[51] Francisco Nunes, Nervo Verdezoto, Geraldine Fitzpatrick, Morten Kyng, Erik Grönvall, and Cristiano Storni. 2015. Self-Care Technologies in HCI: Trends, Tensions, and Opportunities. *ACM Trans. Comput.-Hum. Interact.* 22, 6, Article 33 (2015), 45 pages. https://doi.org/10.1145/2803173

[52] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *PLOS ONE* 15 (2020), e0229132. https://doi.org/10.1371/journal.pone.0229132

[53] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Elizabeth Kaziunas, Stina Matthiesen, and Farah Magrabi. 2021. Realizing AI in Healthcare: Challenges Appearing in the Wild. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21).* Association for Computing Machinery, New York, NY, USA, Article 108, 5 pages. https://doi.org/10.1145/3411763.3441347

[54] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the Impact of Explanations on Advice-Taking: A User Study for AI-Based Clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22).* Association for Computing Machinery, New York, NY, USA, Article 568, 9 pages. https://doi.org/10.1145/3491102.3502104

[55] Aditya kumar Purohit, Aditya Upadhyaya, and Adrian Holzer. 2023. ChatGPT in Healthcare: Exploring AI Chatbot for Spontaneous Word Retrieval in Aphasia. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '23 Companion).* Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3584931.3606993

[56] Emre Sezgin. 2023. Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. *Digit. Health* 9 (Jan. 2023), 20552076231186520.

[57] N. Sadat Shami, Kate Ehrlich, Geri Gay, and Jeffrey T. Hancock. 2009. Making Sense of Strangers' Expertise from Signals in Digital Artifacts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09).* Association for Computing Machinery, New York, NY, USA, 69–78. https://doi.org/10.1145/1518701.1518713

[58] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23).* Association for Computing Machinery, New York, NY, USA, Article 754, 18 pages. https://doi.org/10.1145/3544548.3581075

[59] Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2023. The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users. In *Proceedings of the 5th International Conference on Conversational User Interfaces (CUI '23).* Association for Computing Machinery, New York, NY, USA, Article 2, 10 pages. https://doi.org/10.1145/3571884.3597144

[60] Krzysztof Sobczak, Katarzyna Leoniuk, and Agata Janaszczyk. 2018. Delivering bad news: patient's perspective and opinions. *Patient Prefer. Adherence* 12 (Nov. 2018), 2397–2404.

[61] Michael Spence. 1973. Job Market Signaling. *Q. J. Econ.* 87, 3 (1973), 355. https://doi.org/10.2307/1882010

[62] Michael Spence. 2002. Signaling in retrospect and the informational structure of markets. *Am. Econ. Rev.* 92, 3 (2002), 434–459. https://doi.org/10.1257/00028280260136200

[63] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. The Relationship between Trust in AI and Trustworthy Machine Learning Technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 272–283. https://doi.org/10.1145/3351095.3372834

[64] Niels van Berkel, Maura Bellio, Mikael B. Skov, and Ann Blandford. 2023. Measurements, Algorithms, and Presentations of Reality: Framing Interactions with AI-Enabled Decision Support. *ACM Trans. Comput.-Hum. Interact.* 30, 2, Article 32 (mar 2023), 33 pages. https://doi.org/10.1145/3571815

[65] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (2023), 38 pages. https://doi.org/10.1145/3579605

[66] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 327 (2021), 39 pages. https://doi.org/10.1145/3476068

[67] Joseph B. Walther. 1996. Computer-Mediated Communication: Impersonal, Interpersonal, and Hyperpersonal Interaction. *Communication Research* 23, 1 (1996), 3–43. https://doi.org/10.1177/009365096023001001

[68] Bryan Wang, Gang Li, and Yang Li. 2023. Enabling Conversational Interaction with Mobile UI Using Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 432, 17 pages. https://doi.org/10.1145/3544548.3580895

[69] Mark Warner, Juan F. Maestre, Jo Gibbs, Chia-Fang Chung, and Ann Blandford. 2019. Signal Appropriation of Explicit HIV Status Disclosure Fields in Sex-Social Apps Used by Gay and Bisexual Men. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300922

[70] Dong Wei, Anqi Xu, and Xue Wu. 2020. The mediating effect of trust on the relationship between doctor-patient communication and patients' risk perception during treatment. *PsyCh J.* 9, 3 (June 2020), 383–391.

[71] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 385, 22 pages. https://doi.org/10.1145/3491102.3517582

[72] Ziang Xiao, Q. Vera Liao, Michelle Zhou, Tyrone Grandison, and Yunyao Li. 2023. Powering an AI Chatbot with Expert Sourcing to Support Credible Health Information Access. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. Association for Computing Machinery, New York, NY, USA, 2–18. https://doi.org/10.1145/3581641.3584031

[73] 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*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300468

[74] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. https://doi.org/10.1145/3544548.3581388

[75] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 436, 20 pages. https://doi.org/10.1145/3544548.3581318