

Document Version

Final published version

Licence

CC BY

Citation (APA)

Al Owayyed, M., Brinkman, W. P., Guan, K., Keijsers, L., & Tielman, M. L. (2026). Training Child Helpline Counselors with Value-Integrated Chat Simulations. *International Journal of Human-Computer Interaction*.
<https://doi.org/10.1080/10447318.2026.2645437>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



Training Child Helpline Counselors with Value-Integrated Chat Simulations

Mohammed Al Owayyed, Willem-Paul Brinkman, Kathleen W. Guan, Loes Keijsers & Myrthe L. Tielman

To cite this article: Mohammed Al Owayyed, Willem-Paul Brinkman, Kathleen W. Guan, Loes Keijsers & Myrthe L. Tielman (04 May 2026): Training Child Helpline Counselors with Value-Integrated Chat Simulations, International Journal of Human-Computer Interaction, DOI: [10.1080/10447318.2026.2645437](https://doi.org/10.1080/10447318.2026.2645437)

To link to this article: <https://doi.org/10.1080/10447318.2026.2645437>



© 2026 The Author(s). Published with license by Taylor & Francis Group, LLC



[View supplementary material](#)



Published online: 04 May 2026.



[Submit your article to this journal](#)



Article views: 110








[View related articles](#)



[View Crossmark data](#)

Training Child Helpline Counselors with Value-Integrated Chat Simulations

Mohammed Al Owayyed^{a,b} , Willem-Paul Brinkman^a , Kathleen W. Guan^c , Loes Keijsers^d 
and Myrthe L. Tielman^a 

^aInteractive Intelligence, Delft University of Technology, Delft, The Netherlands; ^bCollege of Computer Science, King Saud University, Riyadh, Saudi Arabia; ^cTechnology, Policy and Management, Delft University of Technology, Delft, The Netherlands; ^dDepartment of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, The Netherlands

ABSTRACT

Children’s helplines train new counselors to adapt to children’s needs and values. This training typically involves roleplay, which can be resource-intensive. Interactive agents offer a promising alternative; yet, simulation-based training systems rarely model how personal values influence decision-making. We present a value-integrated belief–desire–intention (BDI) model that simulates virtual children whose behavior is guided by underlying values. The trainees’ task is to apply motivational interviewing to recognize and align with the child’s values. We conducted a between-subjects experiment ($N = 193$) comparing three conditions: a base BDI virtual child, a BDI virtual child with integrated values, and one with both integrated values and explanatory feedback on value-based reasoning. Results showed credible support that integrating values improves participants’ opportunities to align with a virtual child and enhances their situational awareness based on a child’s values. We also found some support that feedback improved value recognition and perceived usefulness. Additionally, integrating values improved believability and overall experience. These findings suggest that the proposed values-based model enables more targeted training, which we anticipate will better prepare counselors for value-sensitive conversations.

KEYWORDS

BDI; motivational interviewing; child helpline; simulation-based training; value alignment

1. Introduction

When children face challenges such as bullying, emotional distress, or other personal difficulties, one resource they can turn to is a children’s helpline. These helplines offer a confidential and supportive space, typically accessible via a chat-based website or phone, where children can discuss their concerns. However, as the number of annual contacts continues to rise, with a reported increase of over 28% since 2020 (Child Helpline International, 2024), training new counselor volunteers to meet this demand is becoming increasingly difficult. For example, the Dutch Children’s Helpline trained more than 340 new volunteers in 2023, reflecting a 20% increase from the previous year (De Kindertelefoon, 2023). Training typically involves role-playing, with a counselor acting as the child, or supervised sessions with real children led by a trainer. However, this method can be logistically demanding and challenging to scale. One promising solution is interactive agents that can simulate a virtual human in an interactive setting, such as virtual patients (Holderried et al., 2024; Guetterman et al., 2019; Vaughn et al., 2024).

To make such training simulations effective, it is crucial to accurately model the decision-making cues children use during helpline conversations. Human behavior is influenced by many factors, among which are values (Eccles & Wigfield, 2022). Values guide behavior explicitly and implicitly (Keeney, 1988, 1996), shaping how individuals approach interactions and make choices. Like adults, children

CONTACT Mohammed Al Owayyed  M.AlOwayyed@tudelft.nl  Interactive Intelligence, Delft University of Technology, Delft, The Netherlands
 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/10447318.2026.2645437>.

© 2026 The Author(s). Published with license by Taylor & Francis Group, LLC

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

also hold values that inform their decisions (Bubeck & Bilsky, 2004; Döring et al., 2010), such as a preference for “help others” or “be adventurous.” In this paper, we propose enhancing virtual children’s behavior and decision-making by incorporating values into the simulation.

Since children’s values shape their decisions, e.g., whether to share information or withhold it, recognizing and aligning with these values can lead to better counseling. Helpseekers’ values are key in shared decision-making between healthcare providers and those seeking help (Bomhof-Roordink et al., 2019). Counseling textbooks, both general (Committee on Quality of Health Care in America, 2001; McLeod, 2013) and those focused on child helplines (Sindahl, 2013), emphasize the importance of recognizing and aligning with the values of helpseekers. Several prominent therapeutic approaches also highlight the role of values. For instance, client-centred therapy prioritizes the client’s perspective and personal understanding to guide the therapeutic process (Rogers, 1959, 2012). Values clarification, meanwhile, helps clients identify, articulate, and actualize their values to guide behavior and decision-making (Kirschenbaum, 2013). Motivational interviewing (MI) can be seen as a hybrid of client-centred therapy and values clarification (Kirschenbaum, 2013). In MI, the counselor works collaboratively with the client to explore goals and decide on appropriate actions (Miller & Rollnick, 2012). This method has also been used with adolescents (Naar & Suarez, 2021) and in helplines (Pellemans et al., 2024). MI incorporates adolescents’ perspectives to emphasize their autonomy, which is a key aspect of their development (Deci & Ryan, 2013).

Virtual agents have employed MI—either to train counselors through interaction with a virtual patient (e.g., Albright et al., 2018) or to provide counseling via a virtual counselor (e.g., Steenstra et al., 2024). Furthermore, a recent randomized trial found that trainees with limited MI knowledge benefited more from virtual patient simulations than from academic study alone (Zaizar et al., 2024). Despite their importance, values were not explicitly incorporated into human simulations for social skills training. This presents an opportunity to design training agents that help trainees become more aware of and adjust to value-based behaviors by applying MI.

For people to recognize and adapt to values, they need a simulation in which values are integrated, along with an explanation of how these values work. Integrating values in simulations could make virtual interactions more realistic, which in turn may lead to improved training (Saus et al., 2010). Moreover, incorporating values could enable counselors to practice aligning with them and to understand better how values influence a child’s decisions. A necessary first step is learning to recognize which values a child might exhibit. This is similar to recognizing emotions, which can lead to better adjustment (Lawson et al., 2021; Loizou et al., 2024). We adopt a similar line of reasoning for values: it is easier to adjust when one can first identify the underlying value.

However, human values expressed through behavior can be vague (Pincus, 2024), and training with a simulation alone might lack sufficient guidance (Al Owayyed et al., 2025b). Trainees may require explicit guidance through feedback to recognize how a child’s values influence their decisions. Providing feedback on a virtual agent’s decision-making can help interpret its behavior (Harbers et al., 2009), leading to a more guided learning experience (Kirschner et al., 2006). Specifically, feedback on how values shape the child’s decisions may enhance trainees’ pedagogical understanding and support better adjustment.

Previously, we developed Lilobot, a chatbot simulation system designed to train child helpline counselors (Al Owayyed et al., 2024; Grundmann et al., 2025). Lilobot simulates a virtual child contacting a helpline after being bullied at school. The trainee counselor’s task is to follow the five-phase model, a communication protocol commonly used by helplines (Sindahl, 2013). Lilobot is built using a belief–desire–intention (BDI) model (Georgeff et al., 1998). However, values were not yet integrated into Lilobot’s cognition or used to influence its decision-making.

In this paper, we incorporate values into Lilobot’s BDI framework and scenarios, enabling it to simulate a virtual child whose decisions reflect their underlying values. The goal is to train counselors to recognize, understand, and align with children’s values. By explicitly modeling values, we expect it will be easier for trainees to achieve these objectives. Accordingly, this study examines constructs that facilitate learning as reflected in trainees’ interactions with the agent, rather than directly measuring counseling skill acquisition or therapeutic effectiveness.

We therefore examine four constructs. First, recognizing values refers to identifying which values a child expresses. Second, understanding values is captured through situational awareness, reflecting how well trainees understand how a child's values shape their current reasoning and likely actions. Third, we introduce the opportunity to align, describing the extent to which the virtual child's behavior provides clear cues that make value-based alignment possible and actionable. This notion relates to the formation of a therapeutic alliance (Bordin, 1979; Horvath & Greenberg, 1989) and builds on MI, where counselors rely on clients' expressed concerns and values as cues for guiding the conversation. Opportunity to align does not reflect how people learn or acquire skills; rather, it reflects how well the agent is designed to support the required educational objectives. If a training is well-designed, it enables a seamless and easier application of skills (Lee et al., 2020; Merrill, 2012). To illustrate with a driving analogy, this construct reflects how well the crossings and turns are designed in a driving course to teach driving, not the trainee's driving skills themselves. As for the fourth construct, following the Unified Theory of Acceptance and Use of Technology (UTAUT), we assess perceived usefulness as an indicator of trainees' motivation to use the system (Venkatesh et al., 2003). We expect that clearer value cues and alignment opportunities will make the training tool perceived as more useful by supporting task performance. We therefore propose the following hypotheses:

Hypothesis: The inclusion of (a) child values and, additionally, (b) value-related feedback enhances participants' (1) recognition of child values, (2) situational awareness (understanding of the situation), (3) opportunity to align with the virtual child, and (4) perceived usefulness of the training.

To evaluate this approach, we conducted an experiment comparing three conditions: (1) the original BDI model without value modeling, (2) a model with integrated values, and (3) a value-based model accompanied by explanations of the virtual child's reasoning.

2. Related work

2.1. Agents for training helpline counselors

Limited research exists on simulation-based training for helpline counselors, particularly those for children. In the context of suicide helpline training, Demasi et al. (2020) introduced a chatbot designed to simulate a help-seeker. Their system uses machine learning techniques, through annotated transcripts and a multi-task framework, to present various crisis scenarios for training purposes. The authors found that their model improved the diversity and relevance of responses compared to simpler approaches. As for child helplines, the only training simulation we found is a serious game that simulates high-risk scenarios (e.g., trafficking or sexual exploitation) (Veldhuizen-Ochodničanová et al., 2023). This game uses a choice-based interaction model to train counselors in navigating sensitive situations. However, the choice-based design does not allow trainees to practice constructing their own responses, unlike open-ended formats.

2.2. MI agents

Interactive agents have been used to simulate counselors applying MI. In their scoping review, Mercado et al. (2023) identified several cases where conversational agents employed MI to support healthy behaviors, such as addressing alcohol misuse (Rubin et al., 2022), promoting physical activity (Olafsson et al., 2019), and improving women's preconception health (Jack et al., 2020). Agents delivering MI were also shown to improve conversational skills when used before group discussions (Samrose & Hoque, 2022).

On the other hand, interactive agents have also been used to simulate help-seekers, allowing healthcare providers and counselors to train in delivering MI. Results from such training systems are promising. For example, Albright et al. (2018) evaluated a virtual patient training system for substance abuse counseling, finding increased knowledge and skills among trainees compared to a control group. Similarly, Oster et al. (2020, 2025) reported that MI training with a virtual client for diet and weight loss counseling improved

counselors' skills. Other applications include training for HIV care (Rouleau et al., 2020) and clients hesitant to speak with a therapist (Reger et al., 2020).

2.3. Values in social agents

Human values have been modeled in social simulations, which typically involve simulating interactions among multiple agents to study their interactions. For example, Heidari et al. (2020) proposed a value-based framework for simulating multi-agent interactions, applying it to the needs of villagers in a virtual village setting. By incorporating values, the simulation shows how collective behavior can form. Other works also integrate values into multi-agent systems (e.g., van der Weide, 2011; Cointe et al., 2016).

In our case, we simulate values in a single agent interacting with a human user, rather than modeling society-level dynamics. Some frameworks, such as the one proposed by Cranefield et al. (2017), have explored value-based reasoning in BDI agents; however, their approach primarily models the user's values rather than embedding values within the agent itself in a social simulation.

Although we found no agent-based social skills training systems that explicitly model values in their cognitive model, such values are often implicitly present, albeit static and scenario-dependent. For example, in a training system designed to teach de-escalation of aggressive virtual agents through applying specific protocols (Bosse & Provoost, 2014; Bosse et al., 2016), the agent's behavior could be driven by values such as autonomy, respect, or dominance. Similarly, in a vaccination training system (Murali et al., 2022), an anti-vaccination agent may be guided by values like trust, privacy, or safety, leading to preservation of their decision. Values can also influence interactions between a trainee and virtual agents. For instance, when training nurses to speak up to a virtual surgeon (Robb et al., 2015), the doctor's perceived authority might make it harder for nurses to do so. Moreover, agents can reflect different value priorities. In a negotiation training system (Mell & Gratch, 2017), for example, a strategic agent might prioritize achievement, while a free agent might value building trust.

2.4. Children's values

Children's values can affect their decisions and behaviors. They hold such values from an early age, rather than having them emerge only later, although these values are less fine-grained than those held by adults (Uzefovsky et al., 2016; Yalçin, 2021). Even at young ages (e.g., 4–8 years old), children can express value priorities such as benevolence, achievement, security, and autonomy (Scholz-Kuhn et al., 2025; Yalçin, 2021). These value priority categories follow the Schwartz theory of basic human values (Schwartz, 2012), which organizes values into higher-order dimensions. This values model has been applied in developmental research to ground children's value orientations (e.g., Bubeck & Bilsky, 2004; Döring et al., 2010; Scholz-Kuhn et al., 2025). Prior work shows that values can evolve across childhood and adolescence, with different age groups exhibiting different patterns of value change and development. For example, during the early elementary school years, a pattern of increasing self-transcendence (e.g., benevolence) and decreasing self-enhancement (e.g., achievement) is observed over time (Scholz-Kuhn et al., 2025). In this work, we do not aim to model age-specific value development. Instead, we use values that shape children's responses to different situations to model interactions and interpretations. This allows us to focus on how value-sensitive agent behavior can support counselor training through simulations.

3. Integrating values into the BDI model

We have established how incorporating human values into training can benefit the learning process. The next question is: how can such values be integrated into simulation-based training? Specifically, how can we integrate values into the virtual child's BDI model, and how can we provide value-based feedback to explain the child's values and decisions? To answer these questions, we describe Lilobot's system and its components. Furthermore, we illustrate Lilobot's system using a character called Lilo—a virtual child who was bullied at school, values *power* and *achievement* and believes that hitting back will stop the bullying.

3.1. Trainee's learning objective

A trainee needs to understand and align with a child's values to navigate a conversation successfully. To do this, they use MI to guide a child from a potentially conflicting action—such as “hitting back,” which is inconsistent with their value of *security*—toward an alternative action that better aligns with their values—such as “walking away confidently,” which is consistent with both *power* and *security*. This shift leverages cognitive dissonance (Festinger, 1962) in the virtual child's thinking, wherein the child experiences discomfort from the mismatch between their behavior and core values, increasing motivation to adopt actions that better reflect what they care about. Table 3 presents a dialogue example between a trainee and the virtual child, Lilo, illustrating the shift in action.

To facilitate action shift, a trainee should apply the four processes of MI: engaging (building rapport), focusing (identifying the problem), evoking (eliciting change), and planning (developing a plan) (Miller & Rollnick, 2012). To build the system, we used an adolescent-oriented MI guidelines by Naar and Suarez (2021), which incorporates discussions of values across the four MI processes. Specifically, engaging involves clarifying a child's values, focusing entails identifying a discussion path aligned with those values, evoking addresses behavior change in light of the values, and planning ensures the action plan remains consistent with them. These MI processes are applied in the third and fourth phases of the five-phase model (setting a goal with a child and working toward it). In the trainee's intent recognition, we modeled the four processes separately, with each MI process containing 4–9 possible trainee intents. For example, one evoking intent utterance involves addressing the child's action: “What might happen if you chose not to hit back?” By asking this question, the virtual child is triggered to express their reasons for and against the behavior, which supports progress toward change.

3.2. Overall BDI-value model

Before introducing values, we briefly clarify the underlying BDI mechanism used in Lilobot (Grundmann et al., 2025). Lilobot maintains beliefs about itself, the trainee, and the world. It also has desires representing its goals, and intentions representing the actions it is currently committed to. During interaction, user input updates the agent's beliefs, which may lead it to reconsider its current intentions and actions, and to respond accordingly. This allows the agent to simulate how a person reasons about actions over time, rather than simply selecting responses based on predefined outcomes.

Building on this base BDI mechanism, we next describe how values are integrated with the BDI. Figure 1 shows the overall system flow with integrated values, adapted from the ARTES architecture for agent-based social skills training systems (Al Owayyed et al., 2025b). When a trainee inputs a message, the system recognizes it as the trainee's intent and maps it to predefined belief changes within the BDI decision-making framework. For example, if a trainee asks, “What do you think will happen if you hit them back?,” the system identifies this as the trainee's intent to ask Lilo to reflect on its intention (action) of hitting back. Because of this question, Lilo starts to think about their *security*. As a result, Lilo's belief in the BDI that “I think if I hit back I will not be punished” decreases, based on pre-mapping in the reasoning engine. Based on Lilo's updated states, it may maintain or change its desire-intention pair, e.g., shifting from the intention “planning to hit back” to “planning to walk away,”

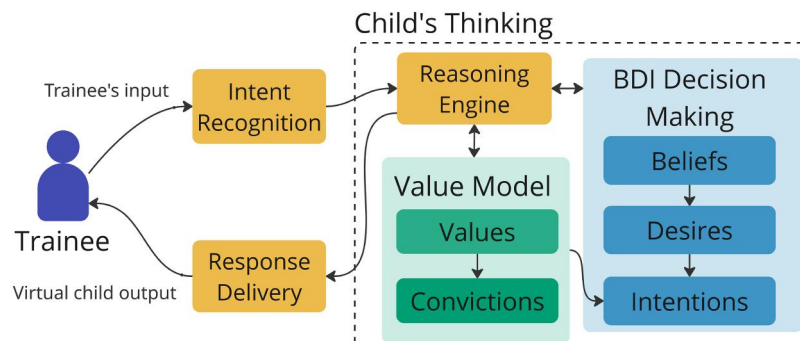


Figure 1. The BDI-values integrated model, based on the ARTES architecture (Al Owayyed et al., 2025b).

where both have the active desire of “Lilo wants to be strong.” Lilo’s response is then selected from pre-defined answers (e.g., “Maybe I’ll get in trouble for fighting”). This response is based on Lilo’s beliefs, active intention, and the trainee’s input. The components in Figure 1 are explained in Table 1. To give an intuition about the value-integrated BDI process, Figure 2 illustrates a single decision cycle of the virtual child, showing how trainee input leads to belief updates and intention selection. The underlying decision-making mechanism is described in detail in Section 3.4.

3.3. Designing value-based virtual children

To create value-based virtual children, we examined literature on values in adolescent bullying. Research by Menesini et al. (2013) shows that *self-enhancement* values (e.g., *power* and *achievement*) positively correlate with bullying, whereas *self-transcendence* values (e.g., *benevolence* and *universalism*) correlate negatively. These value types are defined in Schwartz’s theory of basic human values, in which *self-enhancement* and *self-transcendence* are opposing dimensions (Schwartz, 2012). Therefore, we crafted two persona-based dialogues reflecting these two dimensions, Lilo and Kai. These two virtual children were designed to represent a child in late childhood/early adolescence. Lilo represents self-enhancement, valuing *power* and *achievement*, while Kai represents *self-transcendence*, valuing *benevolence* and *universalism*. Each persona also holds a distinct, yet inconsistent intention regarding their values: Lilo believes hitting back shows strength, which overlooks the possibility that it might lead to punishment; Kai believes keeping the bullying a secret will prevent harm to others, but then the bullying might continue. For the two personas, we defined three core values that fit their scenarios: *self-enhancement*, *self-transcendence*, and *security*—the latter reflecting the shared goal of feeling safe. The personas prioritize the three values differently. Lilo has high *self-enhancement* but low *self-transcendence*, whereas Kai demonstrates the opposite pattern. A full list of Kai and Lilo’s beliefs, desires, intentions, values, and convictions is included in Appendix A.

Table 1. Explanation of components in the value-integrated BDI model.

Component	Explanation
Intent recognition	The system identifies a trainee’s input based on their message (e.g., asking when the bullying occurred), which is then passed to the reasoning engine as a trainee’s intent.
Values (V)	Core principles that guide how a virtual child behaves in a conversation and which actions (intentions) they tend to adopt. Each child is assigned a fixed value (0–1) that shows the strength of the value. We modeled three values: <i>self-transcendence</i> , <i>self-enhancement</i> , and <i>security</i> . These values do not change throughout the interaction.
Convictions (C)	While beliefs reflect a child’s current, changeable perspective on the world, convictions define stable belief–value links that reinforce actions aligned with specific values (e.g., “hitting back leads to power”). Each conviction has a fixed number (0–1) representing the child’s expected impact of an action on a value. Convictions influence the child’s action choices and remain unchanged during a conversation.
Beliefs (B)	Represent what a virtual child currently believes to be true about themselves or the world, along with how strongly these beliefs are held (0–1). These beliefs can be interpreted as statements about the strength of a conviction. Beliefs are dynamic and evolve in response to the trainee’s input.
Desires (D)	Represent the goals the virtual child wants to achieve (e.g., to be strong). These reflect the child’s motivational state.
Intentions (I)	The specific action plans a child is committed to carrying out, based on their current beliefs, desires, convictions, and values (e.g., hitting back or walking away).
Reasoning engine	Manages updates to the BDI components and determines how a child should respond, based on the intent recognition output and the current BDI state. The underlying BDI mechanism without value integration is detailed in Grundmann et al. (2025) and Al Owayyed et al. (2024). Here, the engine also calculates the preferred intention using a weighted combination of belief strength, conviction strength, and value, allowing the system to simulate intention shifts during conversations.
Response delivery	Delivers a virtual child’s responses, as generated by the reasoning engine, to a trainee.

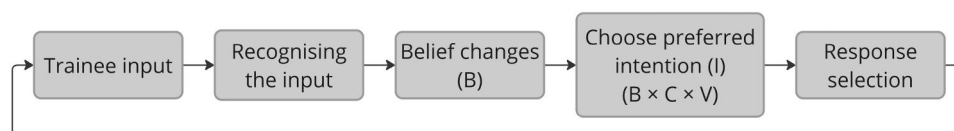


Figure 2. Single decision cycle in the virtual child’s BDI-based reasoning process. Each trainee’s input triggers intent recognition and then belief updates, after which the child reevaluates its intention (action) before generating a response. This cycle is repeated for each trainee input.

3.4. Shifting from inconsistent intention to a consistent one

As a trainee progresses through conversation using MI, a virtual child's inconsistent intention (i.e., planned action) may shift toward a more consistent one aligned with their values. To operationalize the decision-making process, we define the utility of an intention I as a function of the beliefs B_i , convictions C_i , and associated value weights V_i . This produces a utility score per intention:

$$U(I) = \sum_{i=1}^n (B_i \times C_i \times V_i)$$

Given a set of intentions, I , a virtual child selects the intention that maximizes this utility:

$$I^* = \arg \max_{I \in \mathcal{I}} U(I)$$

This mechanism enables a virtual child to dynamically adapt its intention based on changes in beliefs while grounded in a stable set of value-conviction structures. Notably, an agent may select an intention that does not align with its highest-value domain if the supporting beliefs are stronger for an alternative. For example, in Lilo's case, who values *power*, they may choose to walk away rather than retaliate if their belief in punishment, which aligns with *security*, is sufficiently strong. The following are two examples to compute intentions using the numerical values of beliefs, convictions, and values from Table 2:

$$\begin{aligned} \text{Hitting (H)} &= (B_1 \times C_1 \times V_{\text{Power}}) + (B_2 \times C_2 \times V_{\text{Security}}) \\ \text{Hitting (H)} &= (0.9 \times 0.8 \times 0.8) + (0.5 \times 0.5 \times 0.7) \\ &= 0.576 + 0.175 = \mathbf{0.751} \\ \text{Walking away (W)} &= (B_3 \times C_3 \times V_{\text{Security}}) + (B_4 \times C_4 \times V_{\text{Power}}) \\ \text{Walking away (W)} &= (0.4 \times 0.5 \times 0.8) + (0.5 \times 0.6 \times 0.7) \\ &= 0.16 + 0.21 = \mathbf{0.37} \end{aligned}$$

Given the computed utilities, the virtual child selects the intention with the highest utility, which is hitting back¹:

$$I^* = \arg \max_{I \in \{H, W\}} U(I) = \arg \max\{0.751, 0.37\} = H$$

3.5. Illustrative conversational example

Table 3 presents an excerpt of a dialogue between a trainee and Lilo, in which Lilo shifts their intention from "hitting back" to "walking away." The table illustrates how each trainee's input affects the numerical values of relevant beliefs (from Table 2) and the recalculated intention after each belief change. In line 4, the trainee reinforces Lilo's belief that hitting back might make them feel stronger, which increases B1 ("If I hit back, they will think I'm strong"). Then, in line 6, the trainee begins to evoke the motivations behind Lilo's thinking by asking about the possible consequences of hitting back. The system recognizes this as an evocation prompt aimed at encouraging reflection on the inconsistency between Lilo's values and their current intention. As a result, the mapped beliefs are updated—B1 and B2 (beliefs supporting hitting back) decrease. The intention values show that "hitting back" still holds a higher value than "walking away" at this point, so it remains Lilo's preferred intention. In line 8, the

Table 2. Initial beliefs, convictions, and values used to model Lilo's intention preferences. The numbers beside each represent example numerical values used in the dialogue example (Table 3). Lilo also holds additional beliefs, desires, and intentions (Appendix A) that are not relevant to this dialogue example.

Belief statement (B)	Conviction (C)	Value (V)
B1: If I hit back, they will think I'm strong (0.9)	C1: Hitting back → Power (0.8)	Power (0.8)
B2: If I hit back, I will not be punished (0.5)	C2: Hitting back → Security (0.5)	Security (0.7)
B3: If I walk away, they will think I'm strong (0.4)	C3: Walking away → Power (0.5)	
B4: If I walk away, I will not be punished (0.5)	C4: Walking away → Security (0.6)	

Table 3. An excerpt of dialogue between a trainee and Lilo, with changes in beliefs and recalculated intention values. This illustrates the mechanism of belief and preferred intentions shifting in a value-based BDI virtual child.

#	Dialog	Belief change	Intention values	Preferred intention
1	TRAINEE: What do you want to achieve in this conversation?	Initial: B1 = 0.9, B2 = 0.5, B3 = 0.4, B4 = 0.5	Hit = 0.751, Walk = 0.37	Hitting back
2	LILO: <i>I DON'T WANT THEM TO BULLY ME.</i>			
3	LILO: <i>I WANT TO SHOW MY CLASSMATES WHO IS STRONGER.</i>			
4	TRAINEE: Do you feel like hitting back will make you feel stronger?	B1↑ → 1	Hit = 0.815, Walk = 0.37	Hitting back
5	LILO: <i>THEN I WILL HURT THEM BACK AND MAYBE THEY WILL STOP.</i>			
6	TRAINEE: What do you think will happen if you hit them back?	B1↓↓ → 0.8, B2↓↓ → 0.3	Hit = 0.681, Walk = 0.37	Hitting back
7	LILO: <i>MAYBE THE SCHOOL WILL PUNISH ME.</i>			
8	TRAINEE: So you might get in trouble if you fight.	B2↓↓ → 0.1	Hit = 0.611, Walk = 0.37	Hitting back
9	LILO: <i>YESSS.</i>			
10	TRAINEE: What other ways could you show you're strong?	B1↓↓ → 0.6, B2↓ → 0.0, B3↑ → 0.5, B4↑ → 0.7	Hit = 0.512, Walk = 0.494	Hitting back
11	LILO: <i>I CAN WALK AWAY, MAYBE THEY WILL STOP.</i>			
12	TRAINEE: What will happen if you walk away?	B3↑ → 0.6, B4↑ → 0.8	Hit = 0.512, Walk = 0.576	Walking away
13	LILO: <i>THEN THEY MIGHT SAY LILO DOESN'T CARE ABOUT US.</i>			
14	TRAINEE: How do you plan to ignore them when they bully you?	B3↑ → 0.7	Hit = 0.512, Walk = 0.616	Walking away
15	LILO: <i>WHEN THEY TEASE ME, I WILL ACT LIKE THIS DOESN'T BOTHER ME.</i>			
16	TRAINEE: How do you think walking away will make you feel?	B3↑ → 0.8	Hit = 0.512, Walk = 0.656	Walking away
17	LILO: <i>I WILL FEEL BETTER AND STRONGER.</i>			

Note: Arrows indicate the direction of belief change (↑ increase, ↓ decrease, ↓↓ larger decrease); the value after → shows the updated belief strength. The intention column shows the computed utility of each intention, with the preferred intention shown in bold. Appendix E includes conversation scripts of participants interacting with the virtual children during the experiment.

trainee confirms their understanding of Lilo's perspective by asking whether hitting back might lead to trouble. This leads to a further decrease in B2, reflecting Lilo's acknowledgement of possible negative consequences. Lilo confirms this in their response.

From lines 10 to 17, the trainee continues to apply MI techniques, each of which incrementally shifts Lilo's beliefs. These changes gradually alter the calculated utility of the competing intentions. The actual shift in preferred intention occurs at line 12, when the trainee asks what might happen if Lilo walks away. This reinforces beliefs B3 and B4—supporting walking away—and tips the balance in favor of this new intention. In other words, this dialogue sequence illustrates how the child begins to recognize that hitting back may not be consistent with their value of *security*. Ultimately, they prefer to walk away, as it aligns better with both the values of *self-enhancement* and *security* (line 12).

3.6. Feedback generation

The Lilobot system provides feedback to trainees, offering them insight into the virtual child's state. This feedback focuses on how values influence the child's beliefs, decisions, and behavioral shifts throughout the interaction, without explicitly mentioning the values. We modeled two types of feedback based on timing: real-time and post-interaction feedback. Real-time feedback explains the child's current state based on their BDI model and values in real-time. For example, when Lilo has an intention of hitting back, and B4 is less than or equal to B2, trainees receive the following feedback: "The child thinks that hitting back is the only solution that aligns with their values. They have not considered the consequences of such an action." If B4 is higher than B2, but the child still intends to hit back, the feedback is: "The child considered that hitting back might lead to punishment, and they want to avoid that. They seem to be open to considering alternative actions that align with their values." This feedback updates dynamically as the conversation progresses and is generated in real time based on changes in the BDI model. Trainees can access this feedback by clicking a button during the conversation, which opens a pop-up showing explanations of the child's current state (Appendix D).

Post-interaction feedback is presented after the conversation ends, either because the child leaves, the trainee ends the session, or the conversation concludes successfully. This feedback, also based on the final BDI state, explains the child's behavior and decision-making from a value-based perspective. For instance, if the child leaves due to low rapport (i.e., the beliefs related to rapport fall below the threshold required for Lilo to continue the conversation), the feedback states: "At the beginning, the child didn't feel connected enough to open up and discuss solutions. For them, opening up too quickly can feel like giving up control ..."

4. Evaluation

We conducted a between-subjects experiment with three conditions to test the hypotheses presented in Section 1. We randomly assigned participants to one condition: (1) a virtual child with no explicitly modeled values (base BDI; taken from Al Owayyed et al. (2024) and Grundmann et al. (2025)), (2) a virtual child with modeled values (BDI and values), or (3) a virtual child with modeled values and value-based feedback (BDI with values and feedback). Each participant interacted with two virtual child scenarios within the same condition. We asked participants to use MI and the five-phase model to counsel the virtual children. Each experiment took approximately 75 min to complete. Before the experiment, we conducted pilot experiments with 22 participants to identify errors, improve the conversational model, and fine-tune questionnaires. The participants' data were not included in the final analysis.

We obtained ethical approval from TU Delft's Human Ethics Research Committee (HREC reference number: 5191) and preregistered the study on the Open Science Framework.² All data, R analysis code (version 4.3.1), and a Docker container for reproducibility are publicly available via 4TU research data repository.³ The system code is also publicly available on Github.⁴

4.1. Participants

We recruited participants via Prolific from around the world. We required participants to have at least a bachelor's degree (or equivalent), similar to the requirements for volunteering at the Dutch children's helpline. Additionally, participants had to be adults fluent in English with a minimum 95% approval rate and at least five completed studies on Prolific. Although we used Bayesian statistics, we conducted a priori power analysis using frequentist methods to estimate the sample size. Based on a medium effect size, we need 64 participants per condition to achieve 95% confidence and 0.8 power in a *t*-test (Cohen, 1992).

In total, 223 participants started the experiment. Out of those, we excluded 30 participants: 13 failed attention or comprehension checks, 4 restarted the same scenario, 3 experienced technical issues, 5 skipped or wrote very little in the conversation, and 5 appeared to clearly use AI generative tools in the conversation with the virtual child. This left us with 193 participants for the analysis (base condition: $N=64$, values condition: $N=64$, values and feedback condition 3: $N=65$). Table 4 shows a summary of the participant demographics, and Figure 3 presents the average self-reported value profile per condition. Based on these data, we observed no clear differences between the conditions.

4.2. Measures

4.2.1. Primary measures

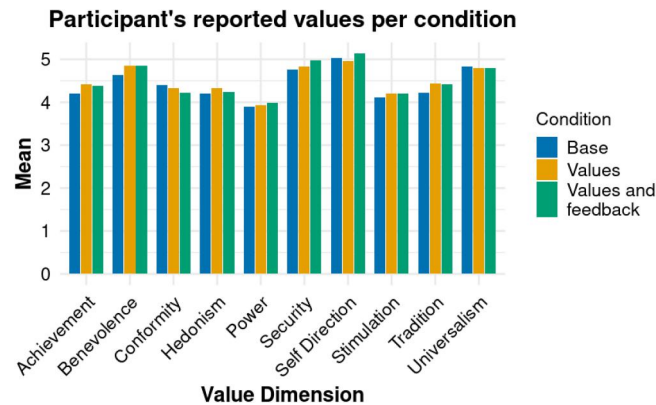
We used four primary measures to answer the hypotheses. All questionnaires are included in Appendix C.

Recognized child values. After each interaction, participants rated the virtual child's values using the Human Values Scale (Schwartz et al., 2015). The scale includes 21 statements aligned with 10 values. Participants rated how much each statement described the child, using a modified version of the original scale: 0 = "Not relevant/Not present/Don't know," 1 = "Not like [child name] at all," up to 6 = "Very much like [child name]." We replaced "me," from the original scale, with either Lilo or Kai, depending on the agent. Furthermore, we calculated the mean of the items per value as recommended

Table 4. Summary of participant demographics by condition (base BDI, BDI with values, BDI with values and feedback).

Demographic	
Gender	
Base	man: 32 (50%), woman: 31 (48.4%), non-binary: 1 (1.6%)
Values	man: 32 (50%), woman: 30 (46.9%), non-binary: 1 (1.6%), prefer not to say: 1 (1.6%)
Values–feedback	man: 38 (58.5%), woman: 26 (40%), prefer not to say: 1 (1.5%)
Age average	
Base	37.42 (13.9)
Values	35.69 (12.7)
Values–feedback	34.97 (12.5)
Education level	
Base	bachelor's: 29 (45.3%), master's: 29 (45.3%), PhD: 4 (6.2%), community college: 2 (3.1%)
Values	bachelor's: 36 (56.2%), master's: 22 (34.4%), PhD: 3 (4.7%), community college: 3 (4.7%)
Values–feedback	bachelor's: 33 (50.8%), master's: 21 (32.3%), PhD: 7 (10.8%), community college: 4 (6.2%)
Country	
Base	South Africa: 23 (35.9%), United States: 19 (29.7%), United Kingdom: 11 (17.2%), others: 11 (17.2%)
Values	South Africa: 24 (37.5%), United States: 15 (23.4%), United Kingdom: 8 (12.5%), others: 17 (26.6%)
Values–feedback	South Africa: 28 (43.1%), United States: 13 (20%), United Kingdom: 11 (16.9%), others: 13 (19.9%)
Has children	
Base	yes: 49 (76.6%), no: 14 (21.9%), prefer not to say: 1 (1.6%)
Values	yes: 45 (70.3%), no: 19 (29.7%)
Values–feedback	yes: 47 (72.3%), no: 18 (27.7%)

Note: All demographic details are available in Appendix B.

**Figure 3.** Participant's average ratings of their own values per condition.

by the scale. To come up with a measure that we can use to test the hypothesis, we calculated the absolute difference in ratings for each value per agent, that is, $|\text{Values}_{\text{Kai}} - \text{Values}_{\text{Lilo}}|$ for each participant. This gave us 10 different scores in values per participant. For additional insights, participants also rated their confidence in their value ratings on a scale from 0 (not confident at all) to 100 (extremely confident).

Situational awareness. We measured participants' understanding of the child's values in a situation using the Situation Awareness Global Assessment Technique (SAGAT) (Endsley, 1988; Sanneman & Shah, 2020). Based on the three levels of situational awareness, we developed four open-ended questions per virtual child: one on perception (level 1), one on comprehension (level 2), and two on projection (level 3), which reflects the highest level of situational awareness. Furthermore, we created a scoring rubric based on how well answers matched the intended child's story details and whether participants linked the child's values to their decisions. We manually score each answer from 0 to 3. We then summed the scores across all four questions to compute a final score out of 12 for each participant, per agent.

Opportunity to align.⁵ We define this as the extent to which participants perceive the virtual child to provide cues that help them align with the child's concerns and values. Initially, we examined existing questionnaires; however, none sufficiently captured this construct. For example, the Working Alliance Inventory (Horvath & Greenberg, 1989) focuses on collaboration toward goals and the relationship between clients and health providers, but it does not assess whether a virtual patient facilitates alignment. We also considered the user-agent alliance construct from the Artificial Social Agent

Questionnaire (ASAQ) (Fitrianie et al., 2025, 2022), but it too emphasizes goal achievement rather than opportunities for value alignment. Therefore, we developed a 7-item scale to measure participants' perceived opportunity to align with and adjust to the child's values. Responses were collected using a 7-point scale ranging from 1 (strongly disagree) to 7 (strongly agree), with intermediate points: 2 (disagree), 3 (somewhat disagree), 4 (neither agree nor disagree), 5 (somewhat agree), and 6 (agree). One example item is: "I adjusted my approach to better align with Kai and Lilo's values."

Perceived usefulness of the training tool. We developed a 7-item questionnaire inspired by the performance expectancy construct from the UTAUT model (Venkatesh et al., 2003), and a prior questionnaire using this construct (Fitrianie et al., 2021). The perceived usefulness questionnaire used the same 7-point scale as the alignment measure and included statements like "The learning tool helps me effectively apply motivational interviewing techniques to align with the virtual child's values."

4.2.2. Secondary measures

To further examine the data, we explored the following secondary measures:

Agent believability. We used the five items in the believability construct from the long ASAQ (Fitrianie et al., 2025, 2022). The scale ranged from -3 (disagree) to $+3$ (agree), with 0 representing "Neither agree nor disagree." In all items, we referred to the agent as the "virtual child."

Overall experience. We measured their experience through the short version of the ASAQ, which has 24 items. Similar to believability, the scale ranged from -3 (disagree) to $+3$ (agree), with 0 representing "Neither agree nor disagree." In all items, we referred to the agent as the "virtual child."

4.3. System

We developed a web-based training platform for the three conditions. The interface displayed a conversational window on the right and two cheat sheets on the left, summarizing the five-phase model and MI. For the BDI with values and value-with-feedback conditions, we used the virtual children model explained in Section 3. For the base BDI condition, we used two virtual children that did not explicitly model values, which were taken from the original system (Al Owayyed et al., 2024; Grundmann et al., 2025). These two virtual children were also involved in bullying and featured two different intentions from each other. The two agents used the same names and age range as the value-based BDI (Lilo and Kai). Visually, the two conditions without feedback (base BDI and BDI with values) had identical interfaces, differing only in the content of the conversation. In the values with feedback condition, the system provided real-time (on-demand via a button) and post-interaction feedback explaining the child's internal value reasoning and state. Screenshots of the system are in Appendix D.

4.4. Procedure

Participants were invited via Prolific and redirected to a Qualtrics survey. There, they first completed an informed consent form and a demographics questionnaire, which included a question about their commitment to providing thoughtful answers in the experiment. After that, they watched two short training videos: one on the five-phase model and the other on MI, focusing on aligning with the child's values. We then randomly assigned participants to one of three conditions. Before starting, participants received a brief explanation of the interface and how to interact with the system. They were then redirected to our system to interact with the first virtual child, where the order of the two virtual children (Kai and Lilo) was randomized. Each conversation had a 13-min limit, during which participants were instructed to continuously interact with the child and restart the conversation if needed. After the time ran out, the participants were returned to Qualtrics to complete two questionnaires: one on situational awareness and one on value recognition, followed by a question on their confidence in the recognized values. Next, participants were redirected to the system for a second 13-min interaction with the other child, followed again by the same set of questionnaires in Qualtrics. Finally, participants completed the remaining measures: the perceived usefulness, the opportunity to align, the short ASAQ, and the believability questionnaires. At the end of the survey, we asked them whether they used a generative AI tool in this experiment and when, and whether they recommend using their data in the study.

4.5. Data analysis

To examine the reliability of the situational awareness scoring, a second coder independently scored a random sample of 19% of all responses. This resulted in a weighted Cohen's kappa of 0.43, which indicates a moderate level of agreement between raters (Landis & Koch, 1977). We also assessed internal consistency for the opportunity to align and perceived usefulness measures; the former showed lower reliability (Cronbach's $\alpha = 0.66$), whereas the latter showed high reliability ($\alpha = 0.92$). For all measures except the child's recognized values, we calculated the mean score per questionnaire and reverse-coded the necessary items, following the guidelines of each scale. We then conducted Bayesian t -tests with uninformative priors to compare the three conditions. For the value recognition measure, we used a multilevel model with uninformative priors, including random intercepts for participants and both random intercepts and slopes by value.

To evaluate the hypotheses, we examined the posterior probabilities and interpreted the strength of evidence based on the guidelines by Chechile (2020) and Andraszewicz et al. (2015). We also inspected the 95% highest density intervals (HDIs) for all measures and reported them to provide additional insight. We conducted the analyses in R (version 4.3.1), using the `BayesFactor` package (Morey & Rouder, 2018) for Bayesian t -tests, the `brms` package (Bürkner, 2017) for multilevel modeling, and the `bayestestR` package (Makowski et al., 2019) to calculate HDIs.

4.6. Results

Here, we report the analysis results of the measures. Table 5 presents the means and standard deviations for all measures across conditions, while Table 6 summarizes the analysis results for the primary and secondary measures. Overall, the results suggest that both the value-integrated and value-with-feedback conditions yielded higher outcomes compared to the base condition.

4.6.1. Primary measures

First, we examine the results for each primary measure, corresponding to the stated hypotheses.

Recognized child values (H1): Participants recognized values most in the values-and-feedback condition. Figure 4 shows the average participant ratings per agent for each condition, where the difference in values across agents is clearer in the two value-based conditions. This is further supported by the standard deviations of value difference scores in Table 5, which are higher in the two value-based conditions. Statistical analysis indicates some support for increased value recognition in the values-

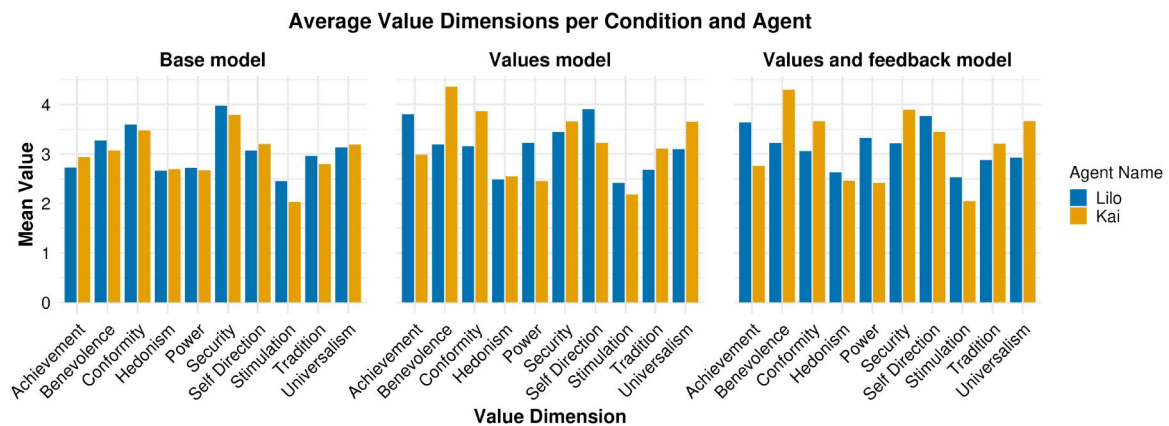
Table 5. Descriptive statistics (means, standard deviations, and 95% highest density intervals [HDIs]) for each condition and hypothesis.

Hypothesis	Condition	Mean (SD)	95% HDI
Recognized child values (H1)	Base	1.03 (0.12)	[0.94, 1.12]
	Values	1.08 (0.23)	[0.92, 1.25]
	Values and feedback	1.18 (0.24)	[1.01, 1.36]
Situational awareness (H2)	Base	4.08 (0.62)	[3.97, 4.19]
	Values	5.38 (1.35)	[5.15, 5.62]
	Values and feedback	5.51 (1.33)	[5.28, 5.74]
Opportunity to align (H3)	Base	4.05 (0.51)	[3.93, 4.18]
	Values	4.41 (0.54)	[4.27, 4.55]
	Values and feedback	4.42 (0.59)	[4.27, 4.57]
Perceived usefulness (H4)	Base	5.82 (0.9)	[5.59, 6.04]
	Values	5.93 (0.87)	[5.71, 6.15]
	Values and feedback	6.01 (0.91)	[5.78, 6.24]
Believability (explorative)	Base	0.94 (1.79)	[0.49, 1.39]
	Values	1.48 (1.48)	[1.11, 1.85]
	Values and feedback	1.34 (1.55)	[0.95, 1.72]
Overall experience (explorative)	Base	0.85 (1.23)	[0.54, 1.15]
	Values	1.19 (1.09)	[0.92, 1.47]
	Values and feedback	1.11 (1.09)	[0.84, 1.38]

Table 6. Summary of the Bayesian test results, including posterior probabilities, 95% highest density intervals (HDIs), and posterior means and standard deviations for the group difference.

Hypothesis	Comparison	Posterior probability	Interpretation	Mean [95% HDI] (SD)
H1: Recognized child values	V > B	0.69	Not worth betting on	0.06 [-0.14, 0.25] (0.12)
	VF > B	0.9	A promising but risky bet	0.16 [-0.04, 0.36] (0.12)
	VF > V	0.8	Only a casual bet	0.10 [-0.10, 0.30] (0.12)
H2: Situational awareness	V > B	>0.99995	Virtually certain	1.29 [1.04, 1.55] (0.13)
	VF > B	>0.99995	Virtually certain	1.42 [1.16, 1.67] (0.13)
	VF > V	0.78	Only a casual bet	0.12 [-0.2, 0.44] (0.16)
H3: Opportunity to align	V > B	0.9996	Nearing certainty	0.35 [0.16, 0.52] (0.09)
	VF > B	0.9999	Nearing certainty	0.36 [0.17, 0.55] (0.10)
	VF > V	0.54	Not worth betting on	0.01 [-0.19, 0.20] (0.10)
H4: Perceived usefulness	V > B	0.76	Only a casual bet	0.11 [-0.20, 0.40] (0.15)
	VF > B	0.87	Only a casual bet	0.18 [-0.12, 0.5] (0.16)
	VF > V	0.69	Not worth betting on	0.08 [-0.23, 0.38] (0.15)
Secondary: believability	V > B	0.97	Good bet—too good to disregard	0.53 [-0.04, 1.08] (0.29)
	VF > B	0.9	A promising but risky bet	0.38 [-0.17, 0.99] (0.29)
	VF > V	0.3	Not worth betting on	-0.14 [-0.66, 0.37] (0.26)
Secondary: Overall experience	V > B	0.95	Good bet—too good to disregard	0.33 [-0.08, 0.72] (0.20)
	VF > B	0.89	Only a casual bet	0.25 [-0.14, 0.66] (0.20)
	VF > V	0.33	Not worth betting on	-0.08 [-0.46, 0.28] (0.19)

Note: B = base BDI condition; V = BDI and values condition; VF = BDI + values + feedback condition. Ha hypotheses (i.e., H1a, H2a, H3a, H4a) refer to the comparison V > B, while Hb hypotheses (i.e., H1b, H2b, H3b, H4b) refer to the comparison VF > V. The interpretations are based on the guidelines by Chechile (2020) and Andraszewicz et al. (2015).

**Figure 4.** How participants rated the values in each agent (Lilo and Kai), across the three conditions.

integrated condition compared to the base condition (H1a: values-integrated > base; posterior probability = 0.69, 95% HDI = [-0.14, 0.25]), and stronger support for the values-and-feedback condition compared to the values-integrated condition (H1b: values-and-feedback > values-integrated; posterior probability = 0.80, 95% HDI = [-0.10, 0.30]). The values-and-feedback condition also shows credible support over the base condition for value recognition (posterior probability = 0.90, 95% HDI = [-0.04, 0.36]).

Confidence ratings showed a similar pattern: participants reported higher confidence in their ratings during the value-integrated condition ($M = 76.7$, $SD = 12.41$) and the values-and-feedback condition ($M = 76.38$, $SD = 17.64$), compared to the base condition ($M = 72.57$, $SD = 19.41$). The posterior probabilities were 0.97 (95% HDI [-0.27, 8.46]) and 0.95 (95% HDI [-0.8, 8.3]), respectively.

Situational awareness (H2): Participants showed higher situational awareness of the child's values in the two values-based conditions (values-only: $M = 5.38$, $SD = 1.35$; values and feedback: $M = 5.51$, $SD = 1.33$) compared to the base condition ($M = 4.08$, $SD = 0.62$). The posterior probabilities provide further credible support that incorporating values improved situational awareness relative to the base condition, with the probability interpreted as virtually certain (H2a; Table 6). Additionally, there is

some support that adding feedback further improved understanding (H2b), with a posterior probability of 0.78 (95% HDI [-0.2, 0.44]).

Opportunity to align (H3): Ratings of alignment opportunity were highest in the values-and-feedback condition ($M = 4.42$, $SD = 0.59$), followed by the values-integrated condition ($M = 4.41$, $SD = 0.54$), and lowest in the base BDI condition ($M = 4.05$, $SD = 0.51$). We also found this pattern in the posterior probability. The probability that the values-integrated condition had a higher rated alignment opportunity than the base condition was 0.9996 (95% HDI [0.16, 0.52]), indicating credible support for H3a. Comparing the values-and-feedback condition to the base condition yielded a posterior probability of 0.9999 (95% HDI [0.17, 0.55]), nearing certainty in favor of the values-and-feedback condition. The posterior probability that the values-and-feedback condition outperformed the values-integrated condition was 0.54 (H3b; 95% HDI [-0.19, 0.20]).

Perceived usefulness (H4): There is some support for H4a and H4b. The values-integrated ($M = 5.93$, $SD = 0.87$) and values-and-feedback ($M = 6.01$, $SD = 0.91$) conditions had posterior probabilities of 0.76 and 0.87, respectively, of being higher than the base BDI condition ($M = 5.82$, $SD = 0.9$). Furthermore, the probability that the values-and-feedback condition had higher perceived usefulness than the values-integrated condition (H4b) was 0.69 (95% HDI [-0.23, 0.38]). Across individual questionnaire items, the values-integrated and values-with-feedback conditions performed similarly, except for an item regarding whether the tool offered insights into the child's reasoning. On this item, the feedback condition scored higher, as expected, since the feedback explicitly articulated the child's state in relation to values (Base-BDI: $M = 5.75$; Values-integrated: $M = 5.73$; Values-and-feedback: $M = 6.06$).

4.6.2. Secondary measures

Like the primary constructs, the secondary measures showed general trends favoring modeled values and feedback.

Believability: Overall, the values-integrated ($M = 1.48$, $SD = 1.48$) and the values-and-feedback ($M = 1.34$, $SD = 1.55$) conditions were rated higher in believability than the base BDI condition ($M = 0.94$, $SD = 1.79$). The posterior probability that the value-integrated virtual child was more believable than the base BDI-only child was 0.97, indicating credible support. The values-and-feedback condition also outperformed the base BDI condition (posterior probability = 0.9). However, the comparison between the two value-modeled conditions (posterior probability = 0.3) provides no support. Thus, adding values to the virtual child simulation appears to improve believability.

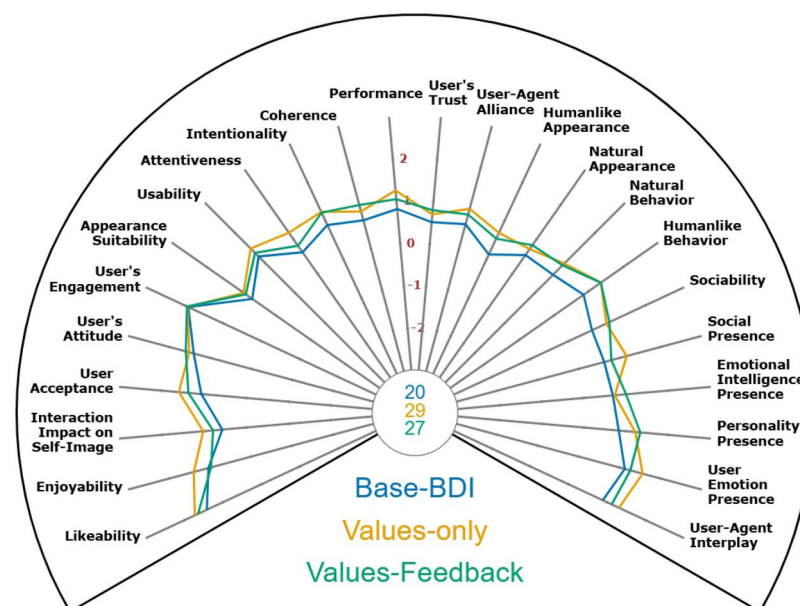


Figure 5. Comparison between the three conditions across all ASAQ constructs. The chart was generated using the official ASAQ visualization tool.

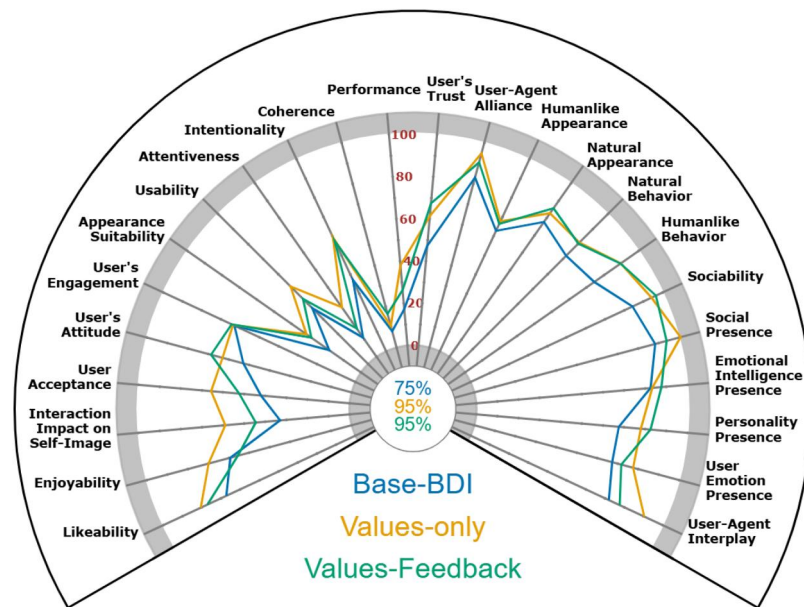


Figure 6. Percentile chart comparing the three agents to the ASAQ Representative Set 2024 (Fitrianie et al., 2025). The gray area indicates the performance range of the 29 agents in the benchmark dataset. The chart was generated using the official ASAQ visualization tool.

Overall experience with the agent: Figure 5 presents a comparison across all ASAQ constructs for the three conditions. The two value-based agents were rated higher than the base condition in nearly all constructs. This is supported by the statistical analysis: the posterior probability that the value-integrated agent outperformed the base condition agent was 0.95, suggesting credible support. The values-and-feedback agent also outperformed the base condition (posterior probability = 0.89), indicating a casual bet. However, the difference between the two value-based agents is not supported (posterior probability = 0.33). To gain a deeper understanding of the experience, we compared our results to those of the ASAQ Representative Set 2024, which comprises a set of 29 artificial social agents that utilized the same questionnaire (Fitrianie et al., 2025). Figure 6 shows that the two value conditions exceed 95% of the set, while the base condition exceeds the 75% mark.

5. Discussion and conclusions

5.1. Results discussion

In the experiment, we examined how integrating values (Ha) and related feedback (Hb) into the simulation-based training system could improve value recognition (H1), situational awareness (H2), opportunity to align (H3), and perceived usefulness of the tool (H4) in a virtual child for the context of training new counselors at helplines.

The results indicate that the training system can help trainees recognize (H1) and understand (H2) the virtual children's values, which could lead to better alignment with those values. This effect was stronger when trainees got feedback. Making the values explicit increased explainability and helped them better understand the virtual child's motivations. These findings support the need to explicitly model values in cognitive models, such as BDI, where they can help explain internal states rather than relying on a black-box model, for example, using a large language model (LLM) in isolation without specifying an internal states model. Similar results have been observed in human-agent teamwork, where people showed higher situational awareness when interacting with explainable agents—those that reveal and explain their beliefs, actions, and decisions (Verhagen et al., 2022).

Generally, participants' identified values matched with the intended values we modeled in the virtual children (Lilo: *power/achievement*; Kai: *benevolence/universalism*), suggesting that participants noticed the intended value cues. An exception was *self-direction*, which was the most frequently identified value

for Lilo, despite not being explicitly intended. One possible explanation is that Lilo's desire to hit the bullies back led participants to perceive Lilo as independent and self-driven. Additionally, the value of *power* may have been under-identified. One reason could be that a related questionnaire item (asking whether "getting rich and expensive stuff" is relevant to Lilo) may not have resonated with the way we expressed *power* through Lilo's actions. This pattern aligns with previous research, which shows that people can instantiate and interpret abstract values in diverse ways (Belic et al., 2022; Hanel et al., 2018).

Results also suggest that integrating values improved the opportunity to align (H3) and the perceived usefulness (H4). Although we did not directly measure learning effects, these results suggest an increase in users' motivation to use the system, which could lead to improved learning outcomes (Lin et al., 2017). In e-learning systems, users' motivation can directly shape how they engage with and accept a system (Esteban-Millat et al., 2018). According to the technology acceptance model, three key factors affect people's motivation to use such systems: perceived usefulness, perceived ease of use, and attitude toward using them (Davis, 1985; Granić & Marangunić, 2019). In our study, we observed an increase in perceived usefulness. Additionally, the opportunity to align can be linked to the perceived usefulness: when trainees can easily identify alignment cues with the virtual child's values, the system becomes more useful to use. This suggests that adding values, particularly when combined with feedback, could positively influence trainees' motivation to use and the actual use of the system.

Beyond the hypotheses, we also explored believability and overall experience. For believability, we had mixed expectations. While we argued that values might enhance believability, it is also possible that a child who does not explicitly articulate their decisions with values (i.e., the base BDI virtual child) may appear equally believable. Results showed credible support that adding values increased believability. This aligns with the findings of Bogdanovych et al. (2016), who reported that virtual characters aware of their own state and their environment were perceived as more believable. On the other hand, the posterior probability that adding feedback alongside values (compared to values-only) improved believability appears to slightly favor the opposite. One possible explanation is that the feedback condition made the virtual child's reasoning overly explicit, articulating thoughts that participants might expect a child to struggle to express, thereby slightly reducing believability. A similar pattern emerged in the overall experience measure, where the posterior probability for adding feedback alongside values (compared to values-only) again slightly favored the opposite. Comparing overall experiences using the short-ASAQ percentile ratings (Figure 6), all three agents scored highly on constructs related to sociality, such as natural behavior, user-agent alliance, and social presence. This supports the idea that the agents provided believable and socially engaging interactions, making them potentially useful for practical applications.

5.2. Limitations and future work

One limitation of this study is that we evaluated the agent's ability to make children's values recognizable and to provide opportunities for alignment, rather than assessing whether the training improved counseling skills, such as the ability to apply MI. Evaluating the impact on counseling skills would require more than a single session, as traditional MI training can take several days to complete (Naar & Suarez, 2021). For example, in a non-randomized controlled trial, training counselors in MI at a suicide helpline produced similar outcomes to training on the five-phase model (Janssen et al., 2022). The study suggested that more extensive MI training could further enhance effectiveness—something we also expect to be necessary in this setting. Given the lack of existing simulation-based training systems with explicitly modeled values, this study serves as an initial step in demonstrating the feasibility and potential benefits of incorporating values into simulated training. For future work, domain experts could assess whether participants applied MI correctly by examining the conversations they had with the virtual children—for example, the conversation scripts from participants in Appendix E.

The situational awareness coding achieved moderate agreement ($\kappa = 0.43$), indicating limitations in the measure's validity that should be taken into account when interpreting the findings. Because values are expressed implicitly through language and behavior, their interpretation is inherently subjective (Gould et al., 2023; Nazirova & Borbala, 2024), which may contribute to variability in situational

awareness assessments. Different raters may emphasize different aspects of participants' reasoning, and alternative codings could therefore lead to some variability in scores that should be considered when interpreting the results. Future research could reexamine the rubric, for example, by explicitly distinguishing between value identification, value–decision reasoning, and action projection.

Regarding the opportunity to align, internal consistency fell below the commonly used 0.70 threshold for acceptable reliability (Bland & Altman, 1997; Nunnally & Bernstein, 1994). To examine the robustness of the findings, we removed two items (items 4 and 5; Appendix C), which resulted in acceptable reliability ($\alpha > 0.70$). Analyses using the reduced scale yielded the same findings, indicating the robustness of the results.

Another limitation lies in the rule-based nature of the tool. In this study, we defined an internal model of values and reasoning, which provides a structured and interpretable foundation. However, intent recognition occasionally failed to identify the trainee's intent accurately, which affected the participants' experience. With recent advances in LLMs, we expect that they can improve intent recognition while maintaining controllability through the BDI-value model (Al Owayyed et al., 2025a). Additionally, the experiment was limited to two scenarios focusing on four values from Schwartz's model, which may limit the generalizability of our findings across all values and scenarios. Furthermore, we focused on incorporating values into a virtual child representing late childhood/early adolescence for training purposes. This abstraction provides a simplified representation of child decision-making that facilitates training, rather than a faithful simulation of age-specific value reasoning. As such, we did not examine how participants' interactions may differ based on the perceived age of a virtual child. Future work could examine whether and how participants' interactions differ based on the perceived age of a virtual child.

While this study considered only the virtual child's values, other stakeholders' values also influence training. These include the values of teachers (Fischer & Hänze, 2020; Thomas, 1977), communities (Halstead & Taylor, 2005), and trainees themselves (Patterson, 1989). The latter is particularly relevant in cases of value conflict during counseling, where discrepancies between a counselor's and a client's values may arise. Several frameworks and strategies exist for managing such conflicts (e.g., Farnsworth & Callahan, 2013; Kirschenbaum, 2013), and incorporating them into future training simulations could help trainees avoid value imposition when interacting with value-based agents. Future research could explore how these stakeholder values can be represented and balanced within training simulations, for example, in scenarios involving value conflicts.

5.3. Ethical implications

Including values in training raises ethical considerations that trainees, tool designers, and trainers must be aware of. In our agents, we illustrated how a child with certain values might behave, for instance, a child who values *power* may choose to hit back. However, this does not represent all possible behaviors associated with those values, which could risk reinforcing stereotypes in training systems (Al Owayyed et al., 2025b). Therefore, training should incorporate a variety of perspectives and virtual child behaviors to avoid associating specific values with action. Trainees should also be mindful of the potential for unethical use of MI to influence behavior, and must adhere to ethical guidelines (Miller & Rollnick, 2012; Naar & Suarez, 2021).

Our training simulation prioritizes internal value tradeoffs from the child's perspective, without explicitly making judgments about what is morally right or aligning with societal norms. While, in our case, avoiding punishment can itself be seen as a kind of moral reasoning, the absence of explicit moral evaluation may lead to undesirable outcomes. For example, if a child does not value *security*, they may still think it is acceptable to hit back, as no conflicting values are guiding them otherwise. This raises questions about how to handle simulated agents with non-conforming or socially problematic values. Such agents might have actions that diverge from accepted norms; on the other hand, they could offer valuable training opportunities by exposing counselors to challenging scenarios. At the same time, deciding which values are “good” or “bad” involves many intersecting factors that make moral judgment complicated (Graham et al., 2013). Rather than enforcing a strict moral framework within the

simulation, training could include reflective exercises that help trainees explore how to respond when a child expresses values they do not share, and facilitate discussions on how to address such situations.

5.4. Conclusion

In conclusion, the findings show that integrating values into a BDI-based conversational agent not only enhances the believability and value recognition of virtual children but also provides trainees with clearer cues for alignment. The value-based model we introduced enables the intentional design of diverse value profiles, allowing for more targeted training to better prepare counselors for value-sensitive conversations.

Notes

1. Where *H* = Hitting and *W* = Walking Away.
2. <https://osf.io/m7x6h>
3. <https://data.4tu.nl/datasets/036a49b4-94f4-4c72-93bd-a874a7172cd6>
4. <https://github.com/alowayyedm/LilobotTraining/tree/values-integrated>
5. In the preregistration, this measure was referred to as “ability to align.” We adopted the term “Opportunity to Align” here to better reflect what it actually measures. The questionnaire items remain the same as those listed in the preregistration.

Acknowledgments

We would like to thank Annemiek Harder for the helpful input and review of the MI conversation questions. We also thank Michaël Grauwde for helping in the analysis, and Carolina Jorge and Nele Albers for their valuable feedback on earlier drafts of this manuscript. We further thank the PROTECTME Consortium and De Kindertelefoon for their helpful input and support.

Author contributions

CRediT: **Mohammed Al Owayyed**: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Writing – original draft, Writing – review & editing; **Willem-Paul Brinkman**: Conceptualization, Formal analysis, Methodology, Project administration, Resources, Writing – review & editing; **Kathleen W. Guan**: Methodology, Resources, Writing – review & editing; **Loes Keijsers**: Methodology, Resources, Writing – review & editing; **Myrthe L. Tielman**: Conceptualization, Formal analysis, Methodology, Project administration, Resources, Writing – review & editing.

Use of generative AI tools

During the preparation of this work, generative AI tools (ChatGPT and Grammarly) were used to enhance language and grammar only. All content was subsequently reviewed and edited by the authors.

Disclosure statement

The authors report there are no competing interests to declare.

Funding

This work and the first author’s research are supported by King Saud University and the Saudi Arabian Cultural Mission (SACM).

ORCID

Mohammed Al Owayyed  <http://orcid.org/0000-0002-9680-9204>

Willem-Paul Brinkman  <http://orcid.org/0000-0001-8485-7092>

Kathleen W. Guan  <http://orcid.org/0000-0002-0044-0140>

Loes Keijsers  <http://orcid.org/0000-0001-8580-6000>
 Myrthe L. Tielman  <http://orcid.org/0000-0002-7826-5821>

Data availability statement

The data collected and analyzed during this study, along with the corresponding analysis code, are available in the 4TU.ResearchData repository at: <https://data.4tu.nl/datasets/036a49b4-94f4-4c72-93bd-a874a7172cd6>.

References

- Al Owayyed, M., Despan, A., Tielman, M., & Brinkman, W.-P. (2024). *A cognitive conversational agent for training child helpline volunteers* [Paper presentation]. Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (pp. 1–3). <https://doi.org/10.1145/3652988.3696197>
- Al Owayyed, M., Denga, A., & Brinkman, W.-P. (2025a). *Controlled yet natural: A hybrid BDI-LLM conversational agent for child helpline training*. In *Proceedings of the ACM International Conference on Intelligent Virtual Agents (IVA '25)*, Berlin, Germany, September 2025. ACM. <https://doi.org/10.1145/3717511.3747075>
- Al Owayyed, M., Tielman, M., Hartholt, A., Specht, M., & Brinkman, W.-P. (2025b). Agent-based social skills training systems: the ARTES architecture, interaction characteristics, learning theories and future outlooks. *Behaviour & Information Technology*, 44(9), 1787–1814. <https://doi.org/10.1080/0144929X.2024.2374891>
- Albright, G., Bryan, C., Adam, C., McMillan, J., & Shockley, K. (2018). Using virtual patient simulations to prepare primary health care professionals to conduct substance use and mental health screening and brief intervention. *Journal of the American Psychiatric Nurses Association*, 24(3), 247–259. <https://doi.org/10.1177/1078390317719321>
- Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., & Wagenmakers, E.-J. (2015). An introduction to bayesian hypothesis testing for management research. *Journal of Management*, 41(2), 521–543. <https://doi.org/10.1177/0149206314560412>
- Belic, J., Djordjevic, A., Nikitović, T., & Khaptsova, A. (2022). The diversity of value construal: A constructivist approach to the Schwartz theory of basic values. *Journal of Constructivist Psychology*, 35(4), 1276–1300. <https://doi.org/10.1080/10720537.2021.1965510>
- Bland, J. M., & Altman, D. G. (1997). [Database] Statistics notes: Cronbach's alpha. *BMJ*, 314(7080), 572. <https://doi.org/10.1136/bmj.314.7080.572>
- Bogdanovych, A., Trescak, T., & Simoff, S. (2016). What makes virtual agents believable? *Connection Science*, 28(1), 83–108. <https://doi.org/10.1080/09540091.2015.1130021>
- Bomhof-Roordink, H., Gärtner, F. R., Stiggelbout, A. M., & Pieterse, A. H. (2019). Key components of shared decision making models: A systematic review. *BMJ Open*, 9(12), e031763. <https://doi.org/10.1136/bmjopen-2019-031763>
- Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice*, 16(3), 252–260. <https://doi.org/10.1037/h0085885>
- Bosse, T., & Provoost, S. (2014). *Towards aggression de-escalation training with virtual agents: A computational model* [Paper presentation]. International Conference on Learning and Collaboration Technologies (pp. 375–387). Springer.
- Bosse, T., Gerritsen, C., & de Man, J. (2016). *An intelligent system for aggression de-escalation training* [Paper presentation]. ECAI 2016 (pp. 1805–1811). IOS Press.
- Bubeck, M., & Bilsky, W. (2004). Value structure at an early age. *Swiss Journal of Psychology*, 63(1), 31–41. <https://doi.org/10.1024/1421-0185.63.1.31>
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Chechile, R. A. (2020). *Bayesian statistics for experimental scientists: A general introduction using distribution-free methods*. MIT Press.
- Child Helpline International (2024). Voices of children & young people around the world: Global child helpline data from 2023, URL <https://childhelplineinternational.org/voices-2023-data/>. Accessed: 2025-04-01.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. <https://doi.org/10.1037/0033-2909.112.1.155>
- Cointe, N., Bonnet, G., & Boissier, O. (2016). *Ethical judgment of agents' behaviors in multi-agent systems* [Paper presentation]. AAMAS (pp. 1106–1114).
- Committee on Quality of Health Care in America (2001). *Crossing the quality chasm: A new health system for the 21st century*. National Academies Press.
- Cranefield, S., Winikoff, M., Dignum, V., & Dignum, F. (2017). *No pizza for you: Value-based plan selection in BDI agents* [Paper presentation]. IJCAI (pp. 178–184).
- Davis, F. D. (1985). *A technology acceptance model for empirically testing new end-user information systems: Theory and results* [PhD thesis]. Massachusetts Institute of Technology.
- De Kindertelefoon (2023). *Jaarrekening 2023 Stichting de Kindertelefoon* <https://jaarverslag.kindertelefoon.nl/resources/jaarverslagen/Jaarrekening%202023%20Stichting%20de%20Kindertelefoon.pdf>

- Deci, E. L., & Ryan, R. M. (2013). *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media.
- Demasi, O., Yu Li., & Z., Yu. (2020). *A multi-persona chatbot for hotline counselor training* [Paper presentation]. Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 3623–3636). <https://doi.org/10.18653/v1/2020.findings-emnlp.324>
- Döring, A. K., Blauensteiner, A., Aryus, K., Drögekamp, L., & Bilsky, W. (2010). Assessing values at an early age: The picture-based value survey for children (PBVS-C). *Journal of Personality Assessment*, 92(5), 439–448. <https://doi.org/10.1080/00223891.2010.497423>
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53(1), 109–132. <https://doi.org/10.1146/annurev.psych.53.100901.135153>
- Endsley, M. R. (1988). *Situation awareness global assessment technique (SAGAT)* [Paper presentation]. Proceedings of the IEEE 1988 National Aerospace and Electronics Conference (pp. 789–795). IEEE. <https://doi.org/10.1109/NAECON.1988.195097>
- Esteban-Millat, I., Martínez-López, F. J., Pujol-Jover, M., Gázquez-Abad, J. C., & Alegret, A. (2018). An extension of the technology acceptance model for online learning environments. *Interactive Learning Environments*, 26(7), 895–910. <https://doi.org/10.1080/10494820.2017.1421560>
- Farnsworth, J. K., & Callahan, J. L. (2013). A model for addressing client–clinician value conflict. *Training and Education in Professional Psychology*, 7(3), 205–214. <https://doi.org/10.1037/a0032216>
- Festinger, L. (1962). Cognitive dissonance. *Scientific American*, 207(4), 93–102. <https://doi.org/10.1038/scientificamerican1062-93>
- Fischer, E., & Hänze, M. (2020). How do university teachers' values and beliefs affect their teaching? *Educational Psychology*, 40(3), 296–317. <https://doi.org/10.1080/01443410.2019.1675867>
- Fitrianie, S., Bruijnes, M., Abdulrahman, A., & Brinkman, W.-P. (2025). The artificial social agent questionnaire (ASAQ)—Development and evaluation of a validated instrument for capturing human interaction experiences with artificial social agents. *International Journal of Human-Computer Studies*, 199, 103482. <https://doi.org/10.1016/j.ijhcs.2025.103482>
- Fitrianie, S., Bruijnes, M., Li, F., Abdulrahman, A., & Brinkman, W.-P. (2022). The artificial-social-agent questionnaire: Establishing the long and short questionnaire versions. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents* (pp. 1–8).
- Fitrianie, S., Horsch, C., Beun, R. J., Griffioen-Both, F., & Brinkman, W.-P. (2021). Factors affecting user's behavioral intention and use of a mobile-phone-delivered cognitive behavioral therapy for insomnia: A small-scale utaut analysis. *Journal of Medical Systems*, 45(12), 110. <https://doi.org/10.1007/s10916-021-01785-w>
- Georgeff, M., Pell, B., Pollack, M., Tambe, M., & Wooldridge, M. 5th (1998). The belief-desire-intention model of agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages*. Proceedings: *International Workshop, ATAL'98 Paris, France, July 4–75* (pp. 1–10). Springer. https://doi.org/10.1007/3-540-49057-4_1
- Gould, R. K., Soares, T. M., Arias-Arévalo, P., Cantú-Fernandez, M., Baker, D., Eyster, H. N., Kwon, R., Prox, L., Rode, J., Suarez, A., Vatn, A., & Zúñiga-Barragán, J. (2023). The role of value (s) in theories of human behavior. *Current Opinion in Environmental Sustainability*, 64, 101355. <https://doi.org/10.1016/j.cosust.2023.101355>
- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology* (Vol. 47, pp. 55–130). Academic Press.
- Granić, A., & Marangunić, N. (2019). Technology acceptance model in educational context: A systematic literature review. *British Journal of Educational Technology*, 50(5), 2572–2593. <https://doi.org/10.1111/bjet.12864>
- Grundmann, S., Al Owayyed, M., Bruijnes, M., Vroonhof, E., & Brinkman, W.-P. (2025). Lilobot: A cognitive conversational agent to train counsellors at children's helplines: Design and initial evaluation. *Journal of Medical Systems*, 49(1), 5. <https://doi.org/10.1007/s10916-024-02121-8>
- Guetterman, T. C., Sakakibara, R., Baireddy, S., Kron, F. W., Scerbo, M. W., Cleary, J. F., & Fetters, M. D. (2019). Medical students' experiences and outcomes using a virtual human simulation to improve communication skills: Mixed methods study. *Journal of Medical Internet Research*, 21(11), e15459. <https://doi.org/10.2196/15459>
- Halstead, M., & Taylor, M. J. (2005). *Values in education and education in values*. Routledge.
- Hanel, P. H. P., Maio, G. R., Soares, A. K. S., Vione, K. C., de Holanda Coelho, G. L., Gouveia, V. V., Patil, A. C., Kamble, S. V., & Manstead, A. S. R. SR (2018). Cross-cultural differences and similarities in human value instantiation. *Frontiers in Psychology*, 9, 849. <https://doi.org/10.3389/fpsyg.2018.00849>
- Harbers, M., Van Den Bosch, K., & Meyer, J.-J. C. (2009). A study into preferred explanations of virtual agent behavior. In *International Workshop on Intelligent Virtual Agents* (pp. 132–145). Springer. https://doi.org/10.1007/978-3-642-04380-2_17
- Heidari, S., Jensen, M., & Dignum, F. (2020). Simulations with values. In H. Verhagen, M. Borit, G. Bravo, & N. Wijnemans (Eds.), *Advances in social simulation* (pp. 201–215). Springer International Publishing. https://doi.org/10.1007/978-3-030-34127-5_19
- Holderried, F., Stegemann-Philipps, C., Herschbach, L., Moldt, J.-A., Nevins, A., Griewatz, J., Holderried, M., Herrmann-Werner, A., Festl-Wietek, T., & Mahling, M. (2024). A generative pretrained transformer (GPT)-

- powered chatbot as a simulated patient to practice history taking: Prospective, mixed methods study. *JMIR Medical Education*, 10(1), e53961. <https://doi.org/10.2196/53961>
- Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the working alliance inventory. *Journal of Counseling Psychology*, 36(2), 223–233. <https://doi.org/10.1037/0022-0167.36.2.223>
- Jack, B. W., Bickmore, T., Yinusa-Nyahkoon, L., Reichert, M., Julce, C., Sidduri, N., Martin-Howard, J., Zhang, Z., Woodhams, E., Fernandez, J., Loafman, M., & Cabral, H. J. (2020). Improving the health of young African American women in the preconception period using health information technology: A randomised controlled trial. *The Lancet. Digital Health*, 2(9), e475–e485. [https://doi.org/10.1016/S2589-7500\(20\)30189-8](https://doi.org/10.1016/S2589-7500(20)30189-8)
- Janssen, W., van Raak, J., van der Lucht, Y., van Ballegooijen, W., & Mérelle, S. (2022). Can outcomes of a chat-based suicide prevention helpline be improved by training counselors in motivational interviewing? A non-randomized controlled trial. *Frontiers in Digital Health*, 4, 871841. <https://doi.org/10.3389/fdgth.2022.871841>
- Keeney, R. L. (1988). Value-driven expert systems for decision support. *Decision Support Systems*, 4(4), 405–412. [https://doi.org/10.1016/0167-9236\(88\)90003-6](https://doi.org/10.1016/0167-9236(88)90003-6)
- Keeney, R. L. (1996). Value-focused thinking: Identifying decision opportunities and creating alternatives. *European Journal of Operational Research*, 92(3), 537–549. [https://doi.org/10.1016/0377-2217\(96\)00004-5](https://doi.org/10.1016/0377-2217(96)00004-5)
- Kirschenbaum, H. (2013). *Values clarification in counseling and psychotherapy: Practical strategies for individual and group settings*. Oxford University Press.
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Lawson, A. P., Mayer, R. E., Adamo-Villani, N., Benes, B., Lei, X., & Cheng, J. (2021). Do learners recognize and relate to the emotions displayed by virtual instructors? *International Journal of Artificial Intelligence in Education*, 31(1), 134–153. <https://doi.org/10.1007/s40593-021-00238-2>
- Lee, J., Kim, H., Kim, K. H., Jung, D., Jowsey, T., & Webster, C. S. (2020). Effective virtual patient simulators for medical communication training: A systematic review. *Medical Education*, 54(9), 786–795. <https://doi.org/10.1111/medu.14152>
- Lin, M.-H., Chen, H.-C., & Liu, K.-S. (2017). A study of the effects of digital learning on learning motivation and learning outcome. *Eurasia Journal of Mathematics, Science and Technology Education*, 13(7), 3553–3564. <https://doi.org/10.12973/eurasia.2017.00744a>
- Loizou, M., Arnab, S., Lameris, P., Hartley, T., Loizides, F., Kumar, P., & Sumilo, D. (2024). Designing, implementing and testing an intervention of affective intelligent agents in nursing virtual reality teaching simulations—A qualitative study. *Frontiers in Digital Health*, 6, 1307817. <https://doi.org/10.3389/fdgth.2024.1307817>
- Makowski, D., Ben-Shachar, M. S., & Lüdtke, D. (2019). bayestestr: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, 4(40), 1541. <https://doi.org/10.21105/joss.01541>
- McLeod, J. (2013). *An introduction to counselling*. McGraw-Hill Education (UK).
- Mell, J., & Gratch, J. (2017). Grumpy & Pinocchio: Answering human-agent negotiation questions through realistic agent design. In *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems* (pp. 401–409).
- Menesini, E., Nocentini, A., & Camodeca, M. (2013). Morality, values, traditional bullying, and cyberbullying in adolescence. *The British Journal of Developmental Psychology*, 31(Pt 1), 1–14. <https://doi.org/10.1111/j.2044-835X.2011.02066.x>
- Mercado, J., Espinosa-Curiel, I. E., & Martínez-Miranda, J. (2023). Embodied conversational agents providing motivational interviewing to improve health-related behaviors: Scoping review. *Journal of Medical Internet Research*, 25, e52097. <https://doi.org/10.2196/52097>
- Merrill, M. D. (2012). *First principles of instruction*. John Wiley & Sons.
- Miller, W. R., & Rollnick, S. (2012). *Motivational interviewing: Helping people change*. Guilford Press.
- Morey, R. D., Rouder, J. N. (2018). *Bayesfactor: Computation of Bayes Factors for Common Designs* URL <https://CRAN.R-project.org/package=BayesFactor>
- Murali, P., Nouraei, F., Fallah, M., Kearns, A., Rebello, K., O’Leary, T., Perkins, R., Pierre Joseph, N., Dedier, J., Paasche-Orlow, M., et al. (2022). Training lay counselors with virtual agents to promote vaccination. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents* (pp. 1–8).
- Naar, S., & Suarez, M. (2021). *Motivational interviewing with adolescents and young adults*. Guilford Publications.
- Nazirova, Z., & Borbala, S. (2024). Values, attitudes and the behaviour paradigm: A systematic literature review. *Journal of Human Values*, 30(2), 214–239. <https://doi.org/10.1177/09716858241236902>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill.
- Olafsson, S., O’Leary, T., & Bickmore, T. (2019). Coerced change-talk with conversational agents promotes confidence in behavior change. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare* (pp. 31–40).

- Oster, C., Leibbrandt, R., Schoo, A., Lange, B., Litt, J., Morello, A., Antonello, C., Powers, D., Maeder, A., & Lawn, S. (2025). A feasibility study of teaching motivational interviewing in a fully online environment using a virtual client. *International Journal of Health Promotion and Education*, 63(4), 193–204. <https://doi.org/10.1080/14635240.2022.2047095>
- Oster, C., Schoo, A., Litt, J., Morello, A., Leibbrandt, R., Antonello, C., Powers, D., Lange, B., Maeder, A., & Lawn, S. (2020). Supporting workforce practice change: Protocol for a pilot study of a motivational interviewing virtual client software tool for health professionals. *BMJ Open*, 10(2), e033080. <https://doi.org/10.1136/bmjopen-2019-033080>
- Patterson, C. H. (1989). Values in counseling and psychotherapy. *Counseling and Values*, 33(3), 164–176. <https://doi.org/10.1002/j.2161-007X.1989.tb00758.x>
- Pellemans, M., Salmi, S., Mérelle, S., Janssen, W., & van der Mei, R. (2024). Automated behavioral coding to enhance the effectiveness of motivational interviewing in a chat-based suicide prevention helpline: Secondary analysis of a clinical trial. *Journal of Medical Internet Research*, 26, e53562. <https://doi.org/10.2196/53562>
- Pincus, J. D. (2024). Values as motives: Implications for theory, methods, and practice. *Integrative Psychological & Behavioral Science*, 58(4), 1704–1750. <https://doi.org/10.1007/s12124-024-09817-z>
- Reger, G. M., Norr, A. M., Rizzo, A. S., Sylvers, P., Peltan, J., Fischer, D., Trimmer, M., Porter, S., Gant, P., & Baer, J. S. (2020). Virtual standardized patients vs academic training for learning motivational interviewing skills in the US Department of Veterans Affairs and the US Military: A randomized trial. *JAMA Network Open*, 3(10), e2017348–e2017348. <https://doi.org/10.1001/jamanetworkopen.2020.17348>
- Robb, A., White, C., Cordar, A., Wendling, A., Lampotang, S., & Lok, B. (2015). A comparison of speaking up behavior during conflict with real and virtual humans. *Computers in Human Behavior*, 52, 12–21. <https://doi.org/10.1016/j.chb.2015.05.043>
- Rogers, C. (2012). *Client centered therapy*. Hachette UK.
- Rogers, C. R., et al. (1959). *A theory of therapy, personality, and interpersonal relationships: As developed in the client-centered framework* (Vol. 3). McGraw-Hill.
- Rouleau, G., Pelletier, J., Côté, J., Gagnon, M.-P., Martel-Laferrrière, V., Lévesque, R., & Fontaine, G, SimforHealth (2020). Codeveloping a virtual patient simulation to foster nurses' relational skills consistent with motivational interviewing: A situation of antiretroviral therapy nonadherence. *Journal of Medical Internet Research*, 22(7), e18225. <https://doi.org/10.2196/18225>
- Rubin, A., Livingston, N. A., Brady, J., Hocking, E., Bickmore, T., Sawdy, M., Kressin, N., Saitz, R., & Simon, S. (2022). Computerized relational agent to deliver alcohol brief intervention and referral to treatment in primary care: A randomized clinical trial. *Journal of General Internal Medicine*, 37(1), 70–77. <https://doi.org/10.1007/s11606-021-06945-9>
- Samrose, S., & Hoque, E. (2022). Mia: Motivational interviewing agent for improving conversational skills in remote group discussions. In *Proceedings of the ACM on Human-Computer Interaction* (Vol. 6, pp. 1–24).
- Sanneman, L., & Shah, J. A. (2020). A situation awareness-based framework for design and evaluation of explainable AI. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9–13, 2020, Revised Selected Papers 2* (pp. 94–110). Springer. https://doi.org/10.1007/978-3-030-51924-7_6
- Saus, E. R., Johnsen, B. H., & Eid, J. (2010). Perceived learning outcome: The relationship between experience, realism and situation awareness during simulator training. *International Maritime Health*, 62(4), 258–264.
- Scholz-Kuhn, R., Makarova, E., Bardi, A., Litzellachner, L. F., Benish-Weisman, M., & Döring, A. K. (2025). Children's personal values and their behavior in the classroom in the early elementary school years: Mapping longitudinal trajectories. *European Journal of Psychology of Education = Journal Europeen de Psychologie de L'education*, 40(3), 75. <https://doi.org/10.1007/s10212-025-00966-2>
- Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1), 11. <https://doi.org/10.9707/2307-0919.1116>
- Schwartz, S. H., Breyer, B., & Danner, D. (2015). Human values scale (ESS). *Zusammenstellung Sozialwissenschaftlicher Items Und Skalen (ZIS)*, <https://zis.gesis.org/DoiId/zis234>
- Sindahl, T. N. (2013). *Chat counselling for children and youth - A handbook* (p. 141).
- Steenstra, I., Nouraei, F., Arjmand, M., & Bickmore, T. (2024). *Virtual agents for alcohol use counseling: Exploring LLM-powered motivational interviewing* [Paper presentation]. Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (pp. 1–10). <https://doi.org/10.1145/3652988.3673932>
- Thomas, K. W. (1977). Toward multi-dimensional values in teaching: The example of conflict behaviors. *The Academy of Management Review*, 2(3), 484–490. <https://doi.org/10.2307/257704>
- Uzefovsky, F., Döring, A. K., & Knafo-Noam, A. (2016). Values in middle childhood: Social and genetic contributions. *Social Development*, 25(3), 482–502. <https://doi.org/10.1111/sode.12155>
- van der Weide, T. L. (2011). *Arguing to motivate decisions* [PhD thesis]. Utrecht University,
- Vaughn, J., Ford, S. H., Scott, M., Jones, C., & Lewinski, A. (2024). Enhancing healthcare education: Leveraging chatgpt for innovative simulation scenarios. *Clinical Simulation in Nursing*, 87, 101487. <https://doi.org/10.1016/j.ecns.2023.101487>

- Veldhuizen-Ochodničanová, E., Cotton, R., Putney, C., Brennan, M., & Haahr, M. (2023). *The use of serious games in knowledge-and skills-based digital counselling: Applications to trafficking in persons in the war in Ukraine* [Paper presentation]. Joint International Conference on Serious Games (pp. 290–298). Springer.
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425–478. <https://doi.org/10.2307/30036540>
- Verhagen, R. S., Neerincx, M. A., & Tielman, M. L. (2022). The influence of interdependence and a transparent or explainable communication style on human-robot teamwork. *Frontiers in Robotics and AI*, 9, 993997. <https://doi.org/10.3389/frobt.2022.993997>
- Yalçın, V. (2021). [Database] Moral development in early childhood: Benevolence and responsibility in the context of children's perceptions and reflections. *Educational Policy Analysis and Strategic Research*, 16(4), 140–163. <https://doi.org/10.29329/epasr.2021.383.8>
- Zaizar, E. D., Gramlich, M. A., Rizzo, A., Reger, G. M., & Norr, A. M. (2024). Exploration of the impact of baseline clinician learner characteristics on motivational interviewing skill improvement following training with a virtual standardized patient. *Training and Education in Professional Psychology*, 18(4), 378–385. <https://doi.org/10.1037/tep0000490>

About the authors

Mohammed Al Owayyed is a PhD candidate in the Interactive Intelligence Group at Delft University of Technology and a lecturer at King Saud University. His research focuses on HCI, behavior change technologies, conversational agents, and educational systems that support social skills development through interactive training tools.

Willem-Paul Brinkman is a full professor in digital behavior change at Delft University of Technology. His research focuses on human interaction with AI for behavior change, specifically systems in the health or training domain that incorporate psychology-informed AI, offering psychological assessment, support, therapy, or training.

Kathleen W. Guan is a PhD candidate at Delft University of Technology. Her research examines participatory mechanisms in adolescent precision digital health, with a particular focus on co-creation. She previously completed a Master of Research in Developmental Neuroscience and Psychopathology at University College London and the Yale Child Study Center.

Loes Keijsers is Professor of child and family studies at Erasmus University Rotterdam. She aims to understand how environmental factors, such as parenting and social media, affect adolescent well-being. She introduces novel technologies, such as Experience Sampling methods, and develops gamified eHealth applications to prevent depressive symptoms.

Myrthe L. Tielman is an assistant professor at the Interactive Intelligence group at Delft University of Technology, the Netherlands. Her research interest lies in how AI can better understand people with the goal of developing truly trustworthy AI. To this end, her focus lies on topics like trust, explainability, and values.