



Delft University of Technology

Document Version

Final published version

Citation (APA)

Al Owayyed, M. (2026). *Interactive Simulation-Based Learning Tools for Training Children's Helpline Counsellors*. [Dissertation (TU Delft), Delft University of Technology]. <https://doi.org/10.4233/uuid:8b58f155-7994-4630-8d21-31b7f9ceb44c>

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

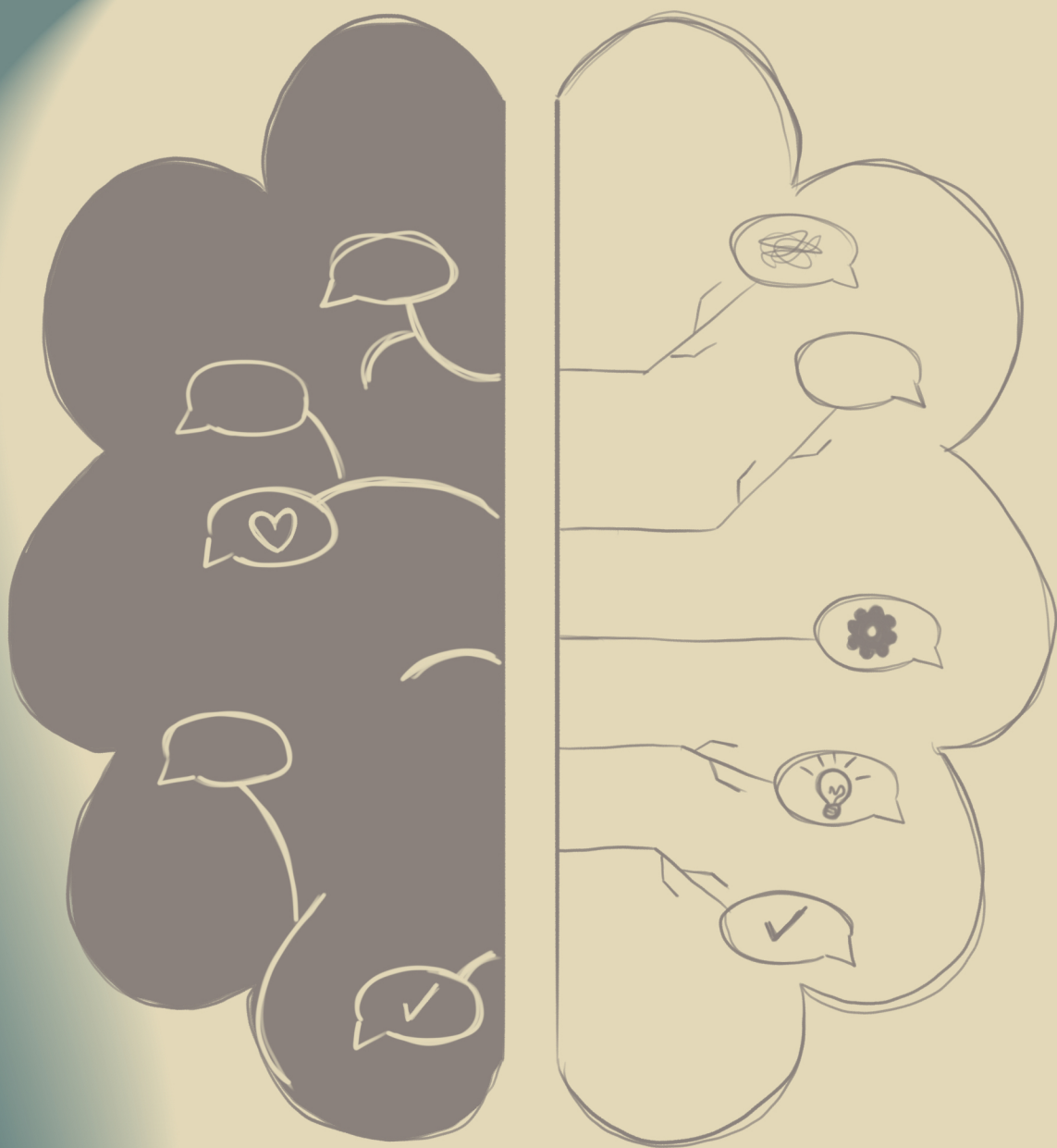
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

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.

This work is downloaded from Delft University of Technology.

Interactive Simulation-Based Learning Tools for Training Children's Helpline Counsellors



Mohammed Al Owayyed

**INTERACTIVE SIMULATION-BASED LEARNING
TOOLS FOR TRAINING CHILDREN'S HELPLINE
COUNSELLORS**

Mohammed AL OWAYYED

INTERACTIVE SIMULATION-BASED LEARNING TOOLS FOR TRAINING CHILDREN'S HELPLINE COUNSELLORS

Dissertation

for the purpose of attaining the degree of doctor
at Delft University of Technology,
by the authority of Rector Magnificus,
Prof. dr. ir. H. Bijl,
chair of the Board for Doctorates,
to be defended publicly on
Wednesday, 10 June 2026, at 15:00 o'clock

by

Mohammed AL OWAYYED

This dissertation has been approved by the promotor.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof. dr. ir. W.P. Brinkman	Delft University of Technology, <i>promotor</i>
Dr. M.L. Tielman	Delft University of Technology, <i>copromotor</i>

Independent members:

Prof. dr. ir. M.S. Kleinsmann	Delft University of Technology
Prof. dr. T. Bosse	Radboud University
Prof. dr. R. Gilissen	Leiden University / 113
Prof. dr. T.W. Bickmore	Northeastern University, United States of America
Prof. dr. M.A. Neerinx	Delft University of Technology, <i>reserve member</i>

SIKS Dissertation Series No. 2026-26.

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.



Keywords: Simulation-based learning; Social skills training; Children's helplines; Counsellor training; Conversational agents; BDI; Human values; Feedback; RCT; Interpretable AI

Printed by: Proefschriftspecialist | <https://www.proefschriftspecialist.nl/>

Cover by: Hanan Alhumaidan

Style: TU Delft House Style, with modifications by Moritz Beller
<https://github.com/Inventitech/phd-thesis-template>

Copyright © 2026 by Mohammed Al Owayyed

ISBN 978-94-6518-320-6

An electronic version of this dissertation is available at <http://repository.tudelft.nl/>.

CONTENTS

Summary	vii
Samenvatting	ix
ملخص البحث	xiii
1 Introduction	1
1.1 Motivation	2
1.2 Research Questions and Hypotheses	3
1.3 Approach	7
2 Agent-Based Social Skills Training Systems: the ARTES Architecture, Interaction Characteristics, Learning Theories and Future Outlooks	11
2.1 Introduction	12
2.2 ARTES Architecture and Characteristics of Social Skills Training Systems .	15
2.3 Key Social Skills Training Components	20
2.4 Relation With the Examined Systems, Other Architectures and Tools . . .	27
2.5 Further Research Directions	36
2.6 Final Remarks	40
3 Simulation-Based Training for Child Helpline Counsellors: Impact of Simulation, Feedback, and Reflection in a Randomised Controlled Trial	41
3.1 Introduction	42
3.2 Hypotheses and Research Questions	43
3.3 Method	46
3.4 Results	59
3.5 Discussion and Conclusions	65
4 Training Child Helpline Counsellors with Value-Integrated Chat Simulations	69
4.1 Introduction	70
4.2 Related Work	72
4.3 Integrating Values Into the BDI Model	73
4.4 Evaluation	80
4.5 Discussion and Conclusions	90
5 Integrating LLMs with a BDI-based Conversational Agent for Child Helpline Training	95
5.1 Introduction	96
5.2 Related Work	97
5.3 Integrating LLMs with Lilobot	98

5.4	Study 1: Script Evaluation	100
5.5	Study 2: Experiment	102
5.6	Study 3: Expert Interviews	109
5.7	Overall Discussion and Conclusions	111
6	Conclusion	113
6.1	Main Conclusions	113
6.2	Limitations	115
6.3	Future Work	117
6.4	Contributions	119
6.5	Ethical Reflection and Implications	121
6.6	Final Remarks	123
A	Appendix for Chapter 1	125
A.1	Overview of Previous Work on Lilobot	126
B	Appendix for Chapter 2	129
B.1	Further explanations of categorisations	130
C	Appendix for Chapter 3	131
C.1	Participant Demographics	132
C.2	Lilobot Training System Feedback	134
C.3	Questionnaires	135
C.4	Extra Analysis	141
D	Appendix for Chapter 4	143
D.1	BDI and Human Values List	144
D.2	Demographics details	145
D.3	Questionnaire Items and Rubric	146
D.4	System Interface Screenshots	148
D.5	Example Participant Conversation Scripts	149
E	Appendix for Chapter 5	153
E.1	Prompt Used for Script Generation	154
E.2	Dialogue Scripts	154
E.3	Guiding Questions for the Semi-Structured Interviews.	156
	Bibliography	159
	SIKS Dissertations	193
	Acknowledgments	205
	Curriculum Vitæ	209
	List of Publications	211

SUMMARY

Children around the world contact children’s helplines when facing emotional, social, or psychological difficulties. These helplines provide confidential support via phone or text-based conversations, where children can share concerns ranging from everyday worries to serious safety issues. Helplines rely on skilled volunteer counsellors who can empathise, structure conversations, and help children find solutions. These helplines train a large number of volunteer counsellors annually to keep up with the volume of conversations they receive. For example, De Kindertelefoon in the Netherlands handled on average around 900 conversations per day and trained 300 new volunteers in 2024. Traditional training methods, such as role-playing, are valuable but resource-intensive, time-consuming, and dependent on the availability of trainers. To address these challenges, interactive simulation-based agents offer a promising extension to existing training practices by enabling scalable, safe, and consistent training. Such agents can simulate a virtual child with whom trainees can practise counselling skills without involving real children. However, current solutions mainly focus on observable interaction behaviour, while paying less attention to clarifying the motivations underlying the child’s actions.

This thesis examines how interactive simulation-based learning tools can train counsellors at children’s helplines. Specifically, it investigates how simulation-based training that includes a virtual child agent can contribute to improved learning outcomes, realism, and educational value while remaining controllable and interpretable. To achieve this, we model more than surface-level dialogue in interactive agents, i.e., the virtual child. These agents should include representations of internal states and how these states change during interaction. One way to achieve this is through the Belief–Desire–Intention (BDI) model, which conceptualises human cognition: beliefs shape desires and guide actions, allowing cognition to be explicitly simulated. This explicitly simulates the internal states, enabling the training system to show not only what a virtual child says but also why their actions and behaviours change. In this work, we develop a training simulation in which a trainee counsels a virtual child contacting a helpline via a text-based interface. The research empirically evaluates combinations of cognitive modelling of a virtual child with educational design to support guided, interpretable, and effective learning experiences.

We first took a broader look at systems that train social skills through interaction with simulated agents. This resulted in ARTES, a general architecture that illustrates how agent-based social skills training systems are structured and how their components work together. This comprehensive architecture clarifies key components and their roles within training systems. ARTES highlights two core elements: the agent simulation that represents a social situation, and educational components that provide guidance and support. We linked ARTES’s components to four learning theories—behaviourism, cognitivism, social cognitive theory, and constructivism—to clarify how the components contribute to learning outcomes. We assessed ARTES’s generalisability, coverage, and implementability by mapping its components across all 43 examined systems, 8 architectures, and 3 implementation tools.

This work also identifies future research directions, including agent simulation fidelity, educational strategies, knowledge transfer, and ethical implications.

Next, we examined how the incremental addition of key components to a simulation-based training system affects learning outcomes. We first examined the effects of simulation in isolation on learning. We then evaluated the impact of adding guidance, for example, through feedback, to the simulation, and subsequently extending it with structured self-reflection. To study this systematically, we conducted a large randomised controlled trial ($N = 346$), in which these three elements were added incrementally. We measured learning outcome measures, such as performance and knowledge gained, as well as trainees' experience and perceived usefulness. The results showed that simulation improved learning outcomes; adding feedback considerably enhanced knowledge and performance; and adding reflection improved reflective writing skills, but dampened performance improvements compared to simulation with feedback.

Modelling human values, such as security or benevolence, helps to explain the virtual child's motivation, as human values influence decision-making and shape how individuals interpret and respond to situations. In this thesis, we demonstrate this by extending the internal-states BDI model of the virtual child agent to explicitly represent values. These modelled values influence the virtual child's beliefs, intentions, and actions. For example, a virtual child may want to hit back after being bullied because they value power and achievement. We evaluated the extended model in a between-subjects experiment ($N = 193$), comparing three agent conditions: without explicitly modelled values, with values, and with values and value-based feedback. The value integrated agent, especially when combined with feedback, showed better outcomes compared to agents without explicitly modelled values.

Although the training improved learning outcomes, its rule-based approach has limitations in understanding diverse and unexpected trainee inputs. Recent advances in generative AI and large language models offer new opportunities to address these limitations. To this end, we integrated a large language model with the existing cognitive model, allowing the LLM to handle language understanding and generation, while the cognitive model retained control over decision-making and the virtual child's internal state. We evaluated this hybrid approach in three ways. First, a script-based evaluation showed that LLM-generated responses were not worse than human-generated ones. Second, a within-subjects study ($N = 37$) showed improved believability, engagement, attitude, and preference compared to the original rule-based system. Third, interviews with helpline experts provided insights into the benefits and concerns of integrating LLMs into training systems, including teachable moments and variability in challenges. These results demonstrate how controlled yet natural interactions can be achieved when integrating LLMs into sensitive training contexts.

Together, these findings show how interpretable, value-aware AI simulations, when combined with pedagogical guidance, can support effective and responsible social skills training. They highlight the potential of simulation-based training to strengthen education in sensitive contexts and improve learning outcomes.

SAMENVATTING

Kinderen over de hele wereld nemen contact op met kindertelefoonlijnen wanneer ze te maken hebben met emotionele, sociale of psychologische problemen. Deze hulpverleners bieden vertrouwelijke ondersteuning via telefoon- of tekstgebaseerde gesprekken, waarin kinderen zorgen kunnen delen die variëren van alledaagse problemen tot ernstige veiligheidskwesties. De gesprekken vinden plaats met goed getrainde vrijwillige hulpverleners, die empathie tonen, gesprekken in goede banen kunnen leiden en kinderen helpen bij het vinden van oplossingen. Om het grote aantal gesprekken te kunnen verwerken, leiden hulpverleners jaarlijks veel nieuwe vrijwilligers op. Zo voerde De Kindertelefoon in Nederland in 2024 ongeveer 900 gesprekken per dag en werden er 300 nieuwe vrijwilligers getraind. Traditionele trainingmethoden, zoals rollenspellen, zijn waardevol maar vergen veel middelen en tijd en zijn afhankelijk van de beschikbaarheid van trainers. Om deze uitdagingen aan te pakken, vormen interactieve, simulatiegebaseerde *agents* een veelbelovende aanvulling op bestaande trainingspraktijken doordat zij schaalbare, veilige en consistente training mogelijk maken. Met dergelijke *agents* kan een virtueel kind worden gesimuleerd, waarmee trainees hun gespreksvaardigheden kunnen oefenen zonder echte kinderen erbij te hoeven betrekken. Bestaande technologische oplossingen richten zich echter vooral op zichtbaar interactiegedrag en besteden minder aandacht aan de onderliggende motivaties achter het gedrag van een virtueel kind.

Dit proefschrift onderzoekt hoe interactieve, simulatiegebaseerde leermiddelen kunnen worden ingezet om hulpverleners bij kindertelefoonlijnen te trainen. Daarbij staat de vraag centraal hoe trainen met een virtueel kind kan bijdragen aan betere leerresultaten, meer realisme en hogere educatieve waarde, terwijl de training controleerbaar en interpreteerbaar blijft. Om dit te bereiken modelleren we meer dan alleen oppervlakkige dialoog in interactieve *agents*, in dit geval een virtueel kind. Deze *agents* bevatten weergaven van mentale toestanden en beschrijven hoe deze tijdens de interactie veranderen. Een manier om dit te realiseren is via het Belief–Desire–Intention (BDI)-model, dat menselijke cognitie conceptualiseert: overtuigingen beïnvloeden verlangens en sturen gedrag, waardoor cognitieve processen expliciet gesimuleerd kunnen worden. Door deze mentale toestanden expliciet te modelleren, kan het trainingssysteem niet alleen laten zien wat een virtueel kind zegt, maar ook waarom diens gedrag en reacties veranderen. In dit proefschrift ontwikkelen we een trainingssimulatie waarin een trainee via een chat interface een virtueel kind begeleidt dat contact opneemt met een hulpverlener. Het onderzoek evalueert empirisch hoe de combinatie van cognitieve modellen van een virtueel kind en een onderwijskundig ontwerp kunnen bijdragen aan interpreteerbare en effectieve leerervaringen.

Allereerst hebben we een brede analyse uitgevoerd van systemen die sociale vaardigheden trainen via interactie met gesimuleerde agents. Dit resulteerde in ARTES, een algemene architectuur die inzichtelijk maakt hoe agentgebaseerde trainingssystemen voor sociale vaardigheden zijn opgebouwd en hoe hun componenten samenwerken. Deze architectuur verduidelijkt de kerncomponenten van dergelijke systemen en hun rol binnen het leerproces. ARTES onderscheidt twee centrale elementen: de agentsimulatie die een sociale

situatie simuleert en de educatieve componenten die begeleiding en ondersteuning bieden. We hebben de componenten van ARTES aan vier leertheorieën gekoppeld – behaviorisme, cognitivisme, sociaal-cognitieve theorie en constructivisme – om te verduidelijken hoe zij bijdragen aan leerresultaten. De generaliseerbaarheid, volledigheid en implementeerbaarheid van ARTES werden geëvalueerd door de componenten te vergelijken met 43 onderzochte systemen, 8 architecturen en 3 implementatietools. Daarnaast identificeert dit werk richtingen voor toekomstig onderzoek, waaronder de validiteit van agentsimulaties, educatieve strategieën, kennisoverdracht en ethische implicaties.

Vervolgens onderzochten we hoe het stapsgewijs toevoegen van belangrijke componenten aan een simulatiegebaseerd trainingssysteem het leerresultaat beïnvloedt. Eerst bestudeerden we het effect van simulatie op zichzelf. Daarna evalueerden we de impact van het toevoegen van educatieve begeleiding, bijvoorbeeld in de vorm van feedback en vervolgens in de vorm van gestructureerde zelfreflectie. Om dit systematisch te onderzoeken voerden we een grootschalige gerandomiseerde controlled trial uit (N = 346), waarin deze drie elementen stapsgewijs werden toegevoegd. We hebben leeruitkomsten gemeten zoals prestatie en opgedane kennis, evenals de ervaring van trainees en hun perceptie van de bruikbaarheid van de training. De resultaten lieten zien dat simulatie het leren verbeterde; dat het toevoegen van feedback leidde tot aanzienlijke verbeteringen in kennis en prestatie; en dat reflectievaardigheden toenamen door zelfreflectie oefeningen, maar dat dit tegelijkertijd de prestatieverbeteringen ten opzichte van simulatie met feedback enigszins afzwakte.

Het modelleren van menselijke waarden, zoals veiligheid of vriendelijkheid, helpt om de motivatie van het virtuele kind beter te verklaren omdat waarden invloed hebben op besluitvorming en bepalen hoe mensen situaties interpreteren en erop reageren. In dit proefschrift laten we dit zien door het interne BDI-model van het virtuele kind uit te breiden met expliciete weergaves van waarden. Deze gemodelleerde waarden beïnvloeden de overtuigingen, intenties en gedrag van het virtuele kind. Zo kan een virtueel kind na pestgedrag willen terug slaan omdat het waarde hecht aan macht en prestatie. We evalueerden dit uitgebreide model in een between-subject experiment (N = 193), waarin drie *agent*condities werden vergeleken: zonder expliciet gemodelleerde waarden, met waarden en met waarden gecombineerd met op waarden gebaseerde feedback. De agent met geïntegreerde waarden, met name in combinatie met feedback, liet betere resultaten zien dan agents zonder expliciet gemodelleerde waarden.

Hoewel de training de leerresultaten verbeterde, kent de regelgebaseerde aanpak beperkingen in het omgaan met uiteenlopende en onverwachte input van trainees. Recente ontwikkelingen op het gebied van generatieve AI en grote taalmodellen bieden nieuwe mogelijkheden om deze beperkingen te overkomen. Daarom integreerden we een large-language model (LLM) met het bestaande cognitieve model, waarbij het LLM verantwoordelijk is voor taalbegrip en -generatie, terwijl het cognitieve model de controle behoudt over besluitvorming en de mentale toestand van het virtuele kind. We evalueerden deze hybride aanpak op drie manieren. Ten eerste liet een evaluatie van de gegenereerde dialogen zien dat door het taalmodel gegenereerde reacties niet slechter waren dan door mensen geschreven reacties. Ten tweede toonde een within-subjects studie (N = 37) verbeteringen in geloofwaardigheid, betrokkenheid, houding en voorkeur ten opzichte van het oorspronkelijke regelgebaseerde systeem. Ten derde leverden interviews met experts van hulplijnen

inzichten op in zowel de voordelen als de zorgen rondom de integratie van LLM in trainingssystemen, waaronder leerzame momenten en variatie in moeilijkheidsgraad. Deze resultaten laten zien je met LLMs interacties kan bereiken die zowel controleerbaar als natuurlijk zijn binnen de context van gevoelige trainingen.

Samen laten deze bevindingen zien hoe interpreteerbare en waarde-gevoelige AI-simulaties, in combinatie met pedagogische begeleiding, effectieve en verantwoorde training van sociale vaardigheden kunnen ondersteunen. Ze onderstrepen het potentieel van simulatiegebaseerde training in gevoelige contexten die leerresultaten kunnen verbeteren.

ملخص البحث

يتواصل الأطفال حول العالم مع خطوط مساعدة الأطفال عند مواجهة صعوبات عاطفية أو اجتماعية أو نفسية. توفر هذه الخطوط دعماً سريعاً عبر الهاتف أو المحادثات النصية لمناقشة مخاوف تتراوح بين هموم يومية ومشكلات خطيرة تتعلق بالسلامة. تعتمد هذه الخدمات على متطوعين مؤهلين في الإرشاد، قادرين على إظهار التعاطف، وتنظيم الحوار، ومساعدة الأطفال على إيجاد حلول مناسبة. نظراً للحجم الكبير من المحادثات، تقوم خطوط المساعدة بتدريب أعداد كبيرة من المتطوعين سنوياً. فعلى سبيل المثال، تعامل خط مساعدة الأطفال في هولندا De Kindertelefoon مع ما يقارب ٩٠٠ محادثة يومياً، ودرّبت نحو ٣٠٠ متطوع جديد في عام ٢٠٢٤. رغم أن أساليب التدريب التقليدية، مثل لعب الأدوار، فعّالة، إلا أنها تستهلك الكثير من الوقت والموارد وتعتمد على توفر المدربين. لمواجهة هذه التحديات، تُمثل أدوات وبرامج التدريب التفاعلية القائمة على المحاكاة إضافة واعدة للممارسات التدريبية الحالية، إذ تتيح تدريباً قابلاً للتوسع، وأماناً، ومتسقاً. يمكن لهذه الأدوات محاكاة طفل افتراضي يتفاعل معه المتدربون لتطوير مهارات الإرشاد دون إشراك أطفال حقيقيين. على الرغم من ذلك، تركز الحلول الحالية على محاكاة السلوك التفاعلي الظاهر، مع اهتمام أقل بتوضيح الدوافع الكامنة وراء تصرفات الطفل وشرح لماذا يتخذ قرارات معينة.

تبحث هذه الأطروحة في كيفية توظيف أدوات التعلّم التفاعلي القائمة على المحاكاة لتدريب المرشدين في خطوط مساعدة الأطفال. على وجه التحديد، تدرس كيف يمكن للتدريب القائم على المحاكاة، والذي يتضمن وكلاً افتراضياً يحاكي طفلاً، أن يسهم في تحسين نواتج التعلّم، وزيادة الواقعية، وتعزيز القيمة التعليمية، مع الحفاظ على قابلية التحكم والتفسير. لتحقيق ذلك، لا نكتفي بنمذجة الحوار السطحي في الوكلاء التفاعليين لحسب، بل نحاكي أيضاً الحالات الداخلية في ذهن الطفل الافتراضي وكيفية تغييرها أثناء التفاعل. يُعد نموذج «المعتقد-الرغبة-النية» (BDI) أحد الأساليب لتحقيق ذلك، إذ يقدم تصوراً معرفياً للإدراك البشري، حيث تصف المعتقدات فهم الطفل لوضعه، وتشكل الرغبات ما يسعى إليه، بينما تمثل النيات الأفعال التي يختار تنفيذها. محاكاة هذا النموذج في ذهن الطفل يسمح بمحاكاة عملية اتخاذ القرار والإدراك. يسمح هذا النهج للنظام التدريبي بإظهار ليس فقط ما يقوله الطفل الافتراضي، بل أيضاً أسباب تغيير سلوكه وتصرفاته. من خلال هذا العمل، نطوّر محاكاة تدريبية يتفاعل فيها المتدرب مع طفل افتراضي يتواصل مع خط المساعدة عبر واجهة نصية. كما نقيم هذا النظام تجريبياً لدراسة أثر الدمج بين الذهن الافتراضي والتصميم التعليمي في دعم تجارب تعلّم موجهة، وقابلة للتفسير، وفعّالة.

بدأنا أولاً بنظرة شاملة على الأنظمة التي تدرب المهارات الاجتماعية من خلال التفاعل مع وكلاء محاكاة. أسفر ذلك عن تطوير بنية عامة سُمّيناها ARTES، والتي توضح كيفية تنظيم أنظمة تدريب المهارات الاجتماعية المعتمدة على الوكلاء وكيفية تكامل مكوناتها. تبرز هذه البنية عنصرين أساسيين: محاكاة الوكيل التي تمثل الموقف الاجتماعي، والمكونات التعليمية التي تقدم التوجيه والدعم. ربطنا مكونات ARTES بأربع نظريات تعلّم-السلوكية، والمعرفية، ونظرية التعلّم الاجتماعي المعرفي، والبنائية-لتوضيح إسهام كل مكون في تحقيق نواتج التعلّم. كما قيّمنا قابلية تعميم

ARTES وشموليتها وإمكانية تطبيقها من خلال مطابقة مكوناتها مع ٤٣ نظاماً تدريبياً، و٨ بنيات، و٣ أدوات تطوير أنظمة. إضافة إلى ذلك، يحدد هذا العمل اتجاهات بحثية مستقبلية، تشمل دقة محاكاة الوكيل، والاستراتيجيات التعليمية، ونقل المعرفة، والجوانب الأخلاقية.

بعد ذلك، درسنا تأثير الإضافة التدريجية لمكونات رئيسية إلى نظام تدريب قائم على المحاكاة على نواتج التعلم للمتدربين. قننا أولاً بدراسة أثر المحاكاة وحدها، ثم قيّمنا تأثير إضافة التوجيه، مثل التغذية الراجعة، وبعد ذلك إضافة التفكير التأملي الذاتي. لتحليل ذلك بشكل منهجي، أجرينا تجربة عشوائية محكمة واسعة النطاق شملت ٣٤٦ مشاركاً، أُضيفت فيها هذه العناصر الثلاثة تدريجياً. من خلال التجربة، قسنا مؤشرات نواتج التعلم، مثل الأداء والمعرفة المكتسبة، إضافة إلى تجربة المتدربين ومدى الفائدة المتصورة من أداة التدريب. أظهرت النتائج أن المحاكاة حسّنت نواتج التعلم، وأن إضافة التغذية الراجعة للمحاكاة عززت المعرفة والأداء بشكل ملحوظ، في حين حسّنت إضافة التفكير التأملي مهارات الكتابة التأملية، لكنها خفّفت من التحسّن في الأداء مقارنة بالمحاكاة مع التغذية الراجعة. يساعد تمثيل القيم الإنسانية، مثل الشعور بالأمان أو الإحسان، في تفسير دوافع الطفل الافتراضي، إذ تؤثر القيم في عملية اتخاذ القرار وفي كيفية فهم الأفراد للواقف واستجابتهم لها. وفي هذه الأطروحة، نُظهر ذلك من خلال توسيع نموذج الحالات الداخلية القائم على «المعتقد-الرغبة-النية» للطفل الافتراضي ليشمل تمثيلاً صريحاً للقيم، بحيث تؤثر هذه القيم الممثلة في معتقدات الطفل الافتراضي على نيّاته وأفعاله. فعلى سبيل المثال، قد يرغب طفل افتراضي في الرد بالعنف بعد التعرّض للتمنر لأنه يقدر القوة والإنجاز. قيّمنا هذا النموذج الموسّع في تجربة شملت ١٩٣ مشاركاً، حيث تفاعل كل مشارك مع أحد ثلاثة أوضاع للوكلاء (الطفل الافتراضي): دون تمثيل صريح للقيم، ومع تمثيل القيم، ومع تمثيل القيم مقروناً بتغذية راجعة قائمة على القيم. وأظهرت النتائج أن الوكيل الذي يدمج القيم—وخاصة عند دعمه بتغذية راجعة—حقق نتائج أفضل مقارنة بالوكلاء الذين لا يمثلون القيم بشكل صريح.

وعلى الرغم من تحسّن نواتج التعلم، فإن المحاكاة القائمة على القواعد تعاني من محدودية في التعامل مع مدخلات المتدربين المتنوعة وغير المتوقعة. توفر التطورات الحديثة في الذكاء الاصطناعي التوليدي ونماذج اللغة الكبيرة فرصاً جديدة لمعالجة هذه القيود. بناءً على ذلك، دمجنا نموذج لغة كبير مع الذهن الافتراضي للطفل، بحيث يتولى نموذج اللغة فهم اللغة وتوليدها، بينما يحتفظ الذهن الافتراضي بالتحكم في اتخاذ القرار والحالات الداخلية للطفل الافتراضي. قيّمنا هذا النهج المهجين بثلاث طرق: أولاً، أظهرت مقارنة قائمة على سيناريوهات أن المحادثات المؤدّة بواسطة نموذج اللغة لم تكن أسوأ من المحادثات المكتوبة يدوياً. ثانياً، بينت دراسة أُجريت على ٣٧ مشاركاً، قام فيها المشاركون بتجربة النظام المهجين والنظام القائم على القواعد فقط، تحسناً في المصادقية، والانخراط، والمواقف الإيجابية، والتفضيل. ثالثاً، وُقّرت مقابلات مع خبراء خطوط المساعدة روى حول فوائد ومخاوف دمج نماذج اللغة في أنظمة التدريب القائمة على المحاكاة، بما في ذلك فرص التعلم وتنوع التحديات. تظهر هذه النتائج كيف يمكن تحقيق تفاعلات طبيعية ومُتمّح بها في آن واحد عند دمج نماذج اللغة الكبيرة في سياقات تدريبية حساسة.

وبشكل عام، تُظهر هذه الأطروحة كيف يمكن لمحاكاة ذكاء اصطناعي قابلة للتفسير وواعية بالقيم، عند دمجها مع توجيه تربي مناسب، أن تدعم تدريباً فعالاً ومسؤولاً على المهارات الاجتماعية. كما تبرز الإمكانيات الكبيرة للتدريب القائم على المحاكاة في تعزيز التعليم في السياقات الحساسة وتحسين نواتج التعلم.

1

INTRODUCTION

1.1 MOTIVATION

Counsellors who support people with psychological and emotional difficulties require strong social and communicative skills to connect with and support their clients [243, 347]. Effective counselling also depends on understanding how to influence and facilitate changes in clients' cognition, emotions, and behaviour. For instance, counsellors often aim to help clients recognise and reframe unhelpful thoughts, shifting their mental states toward healthier ones. This process is at the core of many evidence-based counselling methods, such as Cognitive Behavioural Therapy [33] and Motivational Interviewing [261].

These core counselling skills are also important in more specialised contexts, e.g., child helplines, which are present in many countries. To give an indication of their number worldwide, Child Helpline International is a global network comprising more than 150 child helpline organisations across over 130 countries and territories [79], including the Dutch De Kindertelefoon [201] and the UK Childline [81]. These helplines provide children with a confidential and easily accessible space to share their thoughts, emotions, and concerns. The conversations can range from simple chats to serious emotional or safety-related issues. Children can contact helplines by phone or through text-based chat to collaboratively find a goal. Unlike traditional counselling, counsellors aim to help children within a single conversation. In practice, counsellors follow structured conversation protocols, such as the Five-Phase Model [344], which guides them from rapport-building to closure. They also require a good foundation in conversational and emotional skills, including active listening, empathy, paraphrasing, reflection, summarisation, and indirect leading techniques [177, 279, 344]. Throughout, counsellors need to adapt their communication to the child's pace while maintaining a welcoming and trusting atmosphere.

Training counsellors at scale is, however, challenging. One reason is the high demand, which is particularly relevant for training counsellors at helplines. For example, De Kindertelefoon received an average of 900 conversations per day in 2024 and trained more than 300 new counsellor volunteers that same year [356]. Traditionally, training new counsellors relies on role-playing, in which one trainee plays the child and another the counsellor, guided by a trainer or supervisor. While role-playing provides valuable experiential learning, it can be resource-intensive, time-consuming, and dependent on availability. As a promising extension, simulation-based training uses computer agents that simulate children's behaviour and conversations [143, 313, 319]. Such systems allow trainees to practise their counselling skills in a safe, controllable, and readily available medium where they can make mistakes, experiment with approaches, and receive feedback without compromising real children.

However, simulations designed to train lay counsellors to change help-seekers' thoughts should go beyond surface-level interaction. In counselling and therapeutic contexts, change is achieved by influencing help-seekers' underlying thoughts, beliefs, and interpretations of their situation. Current simulation-based training systems for helpline counsellors typically do not explicitly represent how a client's underlying thoughts and cognitions change during a conversation (e.g., [99, 382]). As a result, trainees might not observe how or why the virtual child's attitudes change in the dialogue. To train these skills effectively, we argue that simulations should model both the child's utterances and their internal states. Such modelling captures how a child's underlying reasoning and cognition might shift in response to the counsellor's input. Including these states could enhance realism

[235, 329] by reflecting how human thinking changes, and it may support education by making these processes explicit and interpretable for trainees. Moreover, representing internal states in virtual agents, such as beliefs and goals, supports the agent's adaptivity [2, 137], allowing educators to guide interactions toward specific learning objectives while maintaining a coherent scenario flow. It also creates an opportunity to build more meaningful feedback—feedback that not only reflects the trainee's performance but also offers insight into how the performance affected the child's thinking. At the same time, it remains unclear how interactive simulation-based training tools should be designed to effectively support training in this context, given challenges such as deciding which internal states to represent, how to balance realism with educational objectives, how to model agent behaviour in an interpretable and pedagogically meaningful way, and how to design feedback that effectively leverages these internal representations. To investigate these opportunities and challenges, we formulate the following main research question:

Main Research Question

In what ways can interactive simulation-based learning tools enhance the training of counsellors in children's helplines?

1.2 RESEARCH QUESTIONS AND HYPOTHESES

To answer the research question, we first need (a) a clear understanding of what a social skills training system should look like. From this, we can identify the key components necessary for such a system. (b) Using these insights, we then examine the components' impact and explore new directions to improve learning outcomes, educational value, realism, and controllability.

1.2.1 WHAT COULD AN OPTIMAL TRAINING SYSTEM LOOK LIKE

Although our focus is on training systems for lay counsellors at helplines, we first take a broader look at systems that train social skills through interaction with simulated agents, including embodied agents. The main goal is to understand what an optimal social skills training system could look like and identify key components.

Developing training environments is inherently complex, as it requires the effective integration of diverse artificial intelligence technologies with domain-specific knowledge [198]. However, the field can benefit from a comprehensive conceptual architecture to abstract this complexity, and describe how components relate to one another. With such an overview, developers and designers can identify relevant elements to build systems that are understandable, adaptable, and complete. Various architectures have been proposed for social skills training, such as the virtual human architecture [223], FAtiMA Toolkit [248], and intelligent tutoring systems [342]. Yet, these architectures often focus either on simulation or on education, rarely combining both effectively. This highlights the need to study how to systematically integrate realistic agent-based simulations with educational components in a comprehensive architecture. A modular architecture that combines both views can indicate which components require greater emphasis, depending on the training context. Therefore, there is a need for an architectural view that integrates simulation

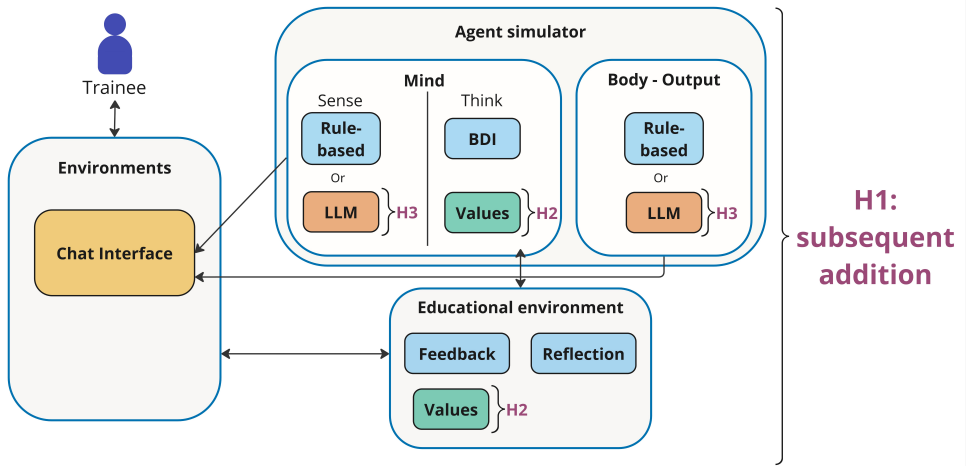


Figure 1.1: Conceptual architecture of a training system based on the ARTES framework. The figure shows how a trainee interacts with the system and highlights the key components examined in this thesis: simulation, feedback, reflection, value integration, and Large Language Models (LLMs). Together, these elements build toward answering the overall research question.

fidelity with educational guidance, while remaining modular enough to adapt to different training contexts.

To address this need and explore how such systems can be structured, we formulate the following research question:

Subquestion 1

What are the key components of simulation-based social skills training systems, and how can they be structured into a comprehensive architecture?

To address the first research question, we analysed existing literature on simulation-based social skills training systems and relevant architectures to identify the key components in such systems. Based on this analysis, we synthesised the components into the ARTES architecture for social skills training systems [8]. In ARTES, these components complement each other: simulations should be realistic while embedding educational elements that enable guided learning. Figure 1.1 presents a simplified model of ARTES and shows how the following hypotheses are structured within it, highlighting the key elements of simulation-based training systems. This includes an agent simulator representing the agent’s decision-making component and an educational environment that provides feedback and guidance to the trainee. For the remainder of this work, we will focus on the components surrounding these two elements.

1.2.2 THE ROLE OF SIMULATION, FEEDBACK AND REFLECTION IN TRAINING SYSTEMS

To examine the impact of key components in simulation-based training systems, we study the combination of simulation, feedback and reflection components. We argue that combining simulation, feedback, and reflection in a training system leads to improved learning outcomes.

To design an effective training system, it is important to know the impact of key elements on learning outcomes. Such elements can be mapped to three primary learning theories that explain how skills are acquired. Behaviourism focuses on shaping behaviour through repetition and reinforcement [345]. Cognitivism emphasises mental processes, viewing learning as the acquisition and structuring of knowledge [113]. Constructivism posits that learners actively build their own understanding through experience [113]. By combining these perspectives, we focus on active, experiential skill acquisition relevant to social skills training.

The first key element in ARTES is simulation. As discussed earlier, simulations can be more cost-effective and practical than traditional training. By using a simulator, trainees interact with representations of real situations, e.g., a virtual child, rather than reading guidelines, enabling experiential learning through repetition and trial-and-error. This aligns with a behaviourist approach [389], where learning occurs through active engagement. However, when learners receive no explicit guidance on why actions are effective or ineffective, this constitutes unguided learning, which has been shown to be less effective than guided learning [85, 204, 366, 368].

To achieve guided learning, feedback can be incorporated into simulations. Feedback could be based on trainees' actions and the virtual child's internal states. By providing such feedback, an otherwise unguided experience is transformed into active thinking about which actions to take and why (aligning with a cognitivist perspective [113]), helping trainees practise effective counselling interactions. Wisniewski et al. [396] conducted a meta-analysis that found that feedback is useful for both cognitive and motivational outcomes. This aligns with Van der Kleij et al. [375], whose review of 40 studies in computer-based environments found that instructional feedback outperforms feedback that merely indicates whether an answer is correct or incorrect. The latter feedback can be conveyed in a simulation (e.g., a virtual child getting frustrated when a trainee makes a mistake), but it fails to provide the explanation required for deep learning. Feedback was also identified as an essential component in social skills training [40]. In addition to cognitive and motivational gains, feedback can also reinforce progress and help trainees monitor their understanding [158].

While feedback provides external guidance, learning can be further strengthened through internal reflection. This aligns with the constructivist perspective [113], where learners actively build their own understanding and interpretation of the world. In the context of social skills training, this means allowing learners to experience scenarios and interpret their own mistakes based on such experience. According to Kolb's experiential learning theory [210], reflection helps trainees consolidate experience into knowledge by connecting what they did, why they did it, and what could be improved. Self-reflection has been found to improve interpersonal skills and knowledge acquisition [168], support knowledge retention [73, 219], as well as lead to higher learning outcomes when combined

with feedback [244].

By combining simulation (experience), feedback (guidance), and reflection (internalisation), we expect to improve learning outcomes. Based on this argument, we formulate the following hypothesis:

Hypothesis 1

Simulation with feedback and reflection in social skills training systems improves learning.

1.2.3 INTEGRATING HUMAN VALUES INTO THE INTERNAL STATES

In this thesis, we argue that integrating human values into a simulation enhances both realism and pedagogical depth, potentially enabling trainees to experience and practise value-sensitive interactions.

Human values—such as security, benevolence, or power—guide behaviour, influence decisions, and determine what individuals consider important [109, 196, 197]. Like adults, children hold values that shape their choices [66, 105], such as whether to seek help, disclose information, or persist in certain actions. Despite their importance, values are rarely modelled explicitly in simulation-based social skills training. Prior work on value modelling has mainly focused on collective behaviour in multi-agent systems rather than on simulating how an individual agent’s values shape decision-making during interaction with a human trainee [89, 161]. As a result, existing training simulations often lack mechanisms that allow trainees to observe, recognise, and respond to value-driven behaviour.

For counsellors, recognising these values is essential, as values often underlie clients’ motivations and choices [324]. This is particularly relevant in frameworks such as Motivational Interviewing (MI) [261], where counsellors aim to understand clients’ perspectives, elicit change, and align interventions with the clients’ values and needs. MI emphasises exploring how certain behaviours (e.g., aggression) may conflict with a person’s beliefs (e.g., wanting to be safe) and supporting them in finding alternative actions that better align with what they care about.

Integrating values into a simulated agent’s internal reasoning may, therefore, enhance both realism and pedagogical relevance. By embedding values within a cognitive architecture such as a BDI model, a virtual agent can exhibit behaviour that reflects priorities while remaining responsive to trainee input. This creates opportunities for trainees to practise identifying value cues and adjusting their counselling strategies accordingly, in line with established counselling frameworks such as MI.

However, recognising and interpreting values can be challenging. Unlike emotions, values are often abstract and expressed indirectly [305]. Without explicit guidance, trainees may overlook how a virtual child’s statements reflect underlying values. Research on virtual agents indicates that explanatory feedback generally helps users interpret decision-making (e.g., [151]). Providing feedback that explains how a simulated agent’s values influence its reasoning may therefore support trainees in recognising values and understanding how their interventions affect the agent’s internal state.

We formulate a hypothesis to test the argument that adding values to the agent’s internal states enhances realism and the trainee’s ability to align with the simulation:

Hypothesis 2

Integrating human values in a simulation improves the learning experience.

1.2.4 INTEGRATING COGNITIVE MODELLING WITH LARGE LANGUAGE MODELS

We take the position that integrating Large Language Models (LLMs) into a structured cognitive model can improve the learning experience in simulation-based training by combining pedagogical control with conversational naturalness.

In simulation-based training, high fidelity to the actual situation is crucial for effective learning [329]. For helpline training, this requires simulating a virtual child capable of realistic, open-ended conversation [38]. However, traditional rule-based agents often struggle to meet this requirement. While they provide a safe, controlled environment, they rely on predefined rules and training data, which limits their ability to process diverse trainee inputs or generate varied responses [311]. Their rigidity can often disrupt conversational flow and undermine realism.

LLMs present a promising solution, as they can enhance intent recognition, provide wider conversational variety, and handle unexpected or nuanced trainee inputs [128, 148, 192]. Yet, relying solely on LLMs for educational simulations introduces concerns of control, predictability, and interpretability. LLMs are not explicitly designed to maintain structured cognitive states or adhere to specific pedagogical constraints. Without such constraints, they may generate unintended content that interferes with learning objectives. Furthermore, LLMs are susceptible to inherent biases, such as formatting and token biases [185, 236]. These reliability issues make pure LLMs risky for sensitive training contexts where specific educational outcomes must be controlled.

A hybrid approach that integrates LLMs within a structured cognitive architecture offers a potential solution to this trade-off. This integration could allow the LLM to handle open-ended dialogue while the BDI cognitive model maintains consistency with the agent's internal states and training goals. This balance aligns with the need for training simulations that are both realistic and pedagogically grounded. Therefore, we test the following hypothesis:

Hypothesis 3

Integrating cognitive models and LLMs in a simulation improves the learning experience.

1.3 APPROACH

Although the hypotheses in this thesis are grounded in existing literature, we conducted empirical research to test them directly. Accordingly, the research approach combines conceptual analysis, system development, and experimental evaluation. We employed a mixed-methods approach to test the hypotheses regarding the system's impact and design. Quantitative measures were used to assess learning outcomes, behaviour, performance,

and interaction perceptions, while qualitative analyses provided insight into participants' reasoning and considerations.

The research follows a systematic approach to operationalise the hypotheses, focusing on a concrete use case to ensure the work remains aligned with practical needs. Furthermore, we repeatedly used the same context to ensure comparability across studies. To this end, we focus on training helpline counsellors using a virtual child simulation, in which the training objective is to change the child's beliefs, thinking, and attitudes. This controlled environment allows us to extend, isolate or adjust specific training variables, such as the presence of feedback or internal values, without the confounding introduced by the unpredictability of real-life counselling.

1.3.1 CONTEXT: LILOBOT - A VIRTUAL CHILD SIMULATION FOR TRAINING CHILDREN'S HELPLINE COUNSELLORS

To ground the research questions and hypotheses in a concrete and realistic training context, this thesis builds on the Lilobot virtual child simulator Grundmann et al. [145]. From the trainee's perspective, Lilobot is a virtual child who has been bullied at school and is contacting a children's helpline for advice via a text-based chat interface. Because this interaction takes place through written chat, Lilobot should realistically represent a child old enough to communicate in writing. A 2025 international child helpline report states that 13-15 year-olds are the most common age group contacting child helplines, while contacts from 10-12 year-olds have increased by more than 50% over the past five years [80]. In the Dutch educational system, children in groups 5 and 6, approximately 8 to 10 years old, are expected to write texts for different purposes while considering audience, structure, and language use [346]. This makes text-based interaction with a children's helpline plausible for children in this age range. According to Piaget's theory of cognitive development, this age range corresponds to the *concrete operational stage*, in which children can reason about concrete events and respond coherently in conversation [121]. Based on these developmental, educational, and helpline-use considerations, Lilobot is best understood as representing a child of at least 9 years old. An initial evaluation of Lilobot showed its potential; however, one limitation was that it was tested with counsellors who already possessed the relevant knowledge, making it not possible to observe how they acquired this information. Therefore, Lilobot should be tested with the intended target group, namely, people without training or experience in helpline counselling.

From a technical perspective, Lilobot is a conversational agent that bases its decision-making on modelled internal states and the trainee's input. These internal states are modelled through the Belief-Desire-Intention (BDI) framework [312]. Lilobot holds beliefs about itself and the surrounding world, such as "I think the trainee is interested in my story" and "I think the trainee can be trusted." The trainee's inputs influence these beliefs—for example, if the trainee asks, "When did the bullying happen?", the aforementioned beliefs are strengthened. Based on its beliefs, Lilobot forms desires (goals it wants to pursue), such as continuing to talk about the bullying if it trusts the trainee enough, or ending the conversation if the trainee has not shown sufficient empathy. During the interaction, the trainee's goal is to apply the Five-Phase communication model correctly. Initial analyses of Lilobot's simulation indicated that the simulation alone may be limited in improving trainees' knowledge and reflective understanding. Further details about the previous work

on Lilobot are available in [145], and an overview is provided in Appendix A.

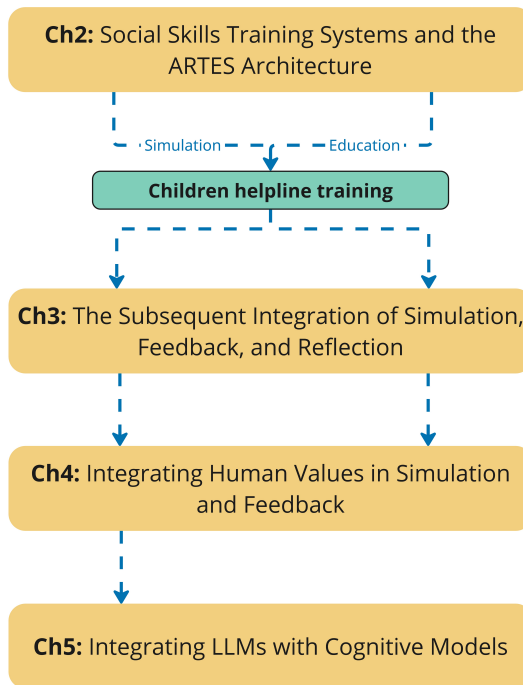


Figure 1.2: Overview of the thesis progression, starting with a conceptual analysis of social skills training systems (RQ1), followed by the context of training new counsellors at child helplines. Within this context, we conducted experimental analyses focusing on both simulation and educational aspects (H1 and H2) or on simulation only (H3).

1.3.2 APPROACH PER RESEARCH QUESTION AND HYPOTHESIS

The overall research flow (Figure 1.2) illustrates the thesis progression, mapping the chapters to the two key elements: simulation and education. The thesis begins more broadly by addressing the need for a clear conceptual understanding of the key components of training systems. To answer this research question, we examined existing social skills training systems and relevant architectures to identify recurring components for inclusion in an architecture, interaction characteristics, and educational elements. Specifically, we analysed eight training system architectures, along with 43 social skills training systems drawn from prior literature reviews, relevant conferences, and targeted searches. This resulted in a modular architecture for social skills training (ARTES). Using ARTES, we identified essential components in training systems, mapped them to existing system designs, and compared ARTES across eight architectures that could be used for building training systems. The resulting architecture guided the design of subsequent experiments. The ARTES architecture and the discussion of social skills training systems are presented in Chapter 2.

To test Hypothesis 1, this thesis evaluates the incremental contribution of simulation, feedback, and reflection to learning outcomes. Lilobot's simulation was extended stepwise with feedback and then reflection, allowing us to systematically study the role of incrementally adding an educational component in supporting guided and effective learning. We conducted a longitudinal randomised controlled trial ($N = 346$), where participants interacted with one version of the tool over five sessions within 14 days. This setup reflects how such training tools could be used in practice, enabling us to observe not only learning outcomes but also how participants' experiences and engagement evolve over repeated interactions. Four additive conditions were compared: a baseline, a simulator-only version, a simulator with feedback, and a simulator with feedback and reflection. We measured changes in learning and performance across the four conditions. More information about the randomised controlled trial setup and results is provided in Chapter 3.

Building on this, we tested whether extending the agent's internal reasoning with human values improves realism and trainee alignment (H2). To simulate how values influence decisions, we extended Lilobot's BDI model with an explicit value layer that shapes the agent's reasoning and behavioural choices. With this, trainees could encounter value-driven behaviour during interaction and practice recognising and aligning with those values, as desirable in real counselling conversations. The experiment compared three conditions in a between-subjects design ($N = 193$): the base Lilobot simulator, the simulator with values, and the simulator with values and feedback. We tested participants' ability to recognise values, perceived alignment, and the extent to which they understand how values influenced the interaction. The BDI-Values model design and experiment are described in Chapter 4.


To evaluate the position that integrating internal states with LLMs improves the experience (H3), we examine how integrating LLMs into a cognitive model affects perceptions of realism, consistency, and suitability for training. In this hybrid BDI-LLM approach, the LLM handled open-ended language understanding and generation while the BDI layer maintained consistency with the agent's internal state and training objectives. We evaluated this approach through three complementary studies: (1) a script-based assessment comparing LLM-generated responses with human-crafted responses to establish whether LLM output meets a minimum quality threshold for training use; (2) a within-subjects experiment ($N = 37$) examining how the hybrid BDI-LLM integration affects trainees' perceptions compared to a rule-based system through direct interaction; and (3) expert interviews to capture concerns about appropriateness and risk that cannot be assessed through participant studies alone. Together, these studies triangulate quantitative user perceptions with qualitative expert insight. Chapter 5 presents the hybrid BDI-LLM integration and evaluation results.

Finally, the concluding chapter (Chapter 6) summarises the findings and draws conclusions across all studies. We highlight the ethical and research implications for designing effective, interpretable, and value-aligned simulation-based training systems. The thesis structure follows a funnel-like progression, from conceptual exploration to detailed experimental testing (Figure 1.2), reflecting how each research question and hypothesis builds upon the previous ones.

2

AGENT-BASED SOCIAL SKILLS TRAINING SYSTEMS: THE ARTES ARCHITECTURE, INTERACTION CHARACTERISTICS, LEARNING THEORIES AND FUTURE OUTLOOKS

Agent-based training systems can enhance people's social skills. The effective development of these systems needs a comprehensive architecture that outlines their components and relationships. Such an architecture can pinpoint improvement areas and future outlooks. This paper presents ARTES: a general architecture illustrating how components of agent-based social training systems work together. We studied existing systems and architectures for training and tutoring to design ARTES and identify its essential components and interaction characteristics. ARTES comprises two core components: the agent simulation of social situations, and educational elements to provide guided learning. We link ARTES's crucial components to four primary learning theories (behaviourism, cognitivism, social cognitive theory, and constructivism) to illustrate the role of agent simulation and tutoring elements in establishing desired learning outcomes. Furthermore, we map ARTES's components against eight architectures, 43 systems and three tools to indicate the components' relevance, completeness, generalisation, and deployment potential across contexts. In addition to ARTES, the paper also contributes by identifying future improvements and research directions, such as the agent's thinking, tutoring methods, knowledge transfer, and ethical implications. We believe ARTES can help bridge the

 **Al Owayyed, M.**, Tielman, M., Hartholt, A., Specht, M., and Brinkman, W. P. (2025). Agent-based social skills training systems: the ARTES architecture, interaction characteristics, learning theories and future outlooks. *Behaviour & Information Technology*, 44(9), 1787-1814.

gap between virtual human simulations and impactful educational learning, offering training system developers desirable features like understandability and adaptability.

2

2.1 INTRODUCTION

Social communication skills are important for human interaction. Training these skills can, therefore, be beneficial, including training them with computer applications. Software engineers designing these systems would benefit from a high-level application overview; that is, a conceptual architecture describing the system's components, their interrelationships, and properties. This overview helps engineers identify what is necessary and build applications with desirable characteristics such as understandability for other developers and adaptability [237]. Nonetheless, the area of social skills training systems currently lacks such an architecture of the components and their relationships. To this end, we present a conceptual architecture for agent-based social skills training systems. This architecture outlines the general functionalities and components necessary to deliver these functionalities. The architecture is versatile and could be applied to various forms of social skills training, including interviews, negotiations, and taking medical histories. The architecture is also agnostic to the agent's embodiment, modality of interactions, software implementation, and deployment. Because of their importance, it caters explicitly for training on dialogues aimed at changing people's perspectives and beliefs, such as getting vaccinated, becoming physically active, or quitting smoking. Therefore, we specifically focus on simulating the thought process of a communication partner but also on how the architecture facilitates the embedding of instructional principles. Moreover, we also look into new opportunities and further research directions for agent-based social skills training.

2.1.1 WHY SOCIAL SKILLS TRAINING SYSTEMS

Social skills training is beneficial to train people on problems arising in different social settings, such as social anxiety [35], patient-doctor communication [288], and business customer service [269]. Yet, this training is often reported to be costly, time-consuming, and involves several individuals, e.g., training student doctors on breaking bad news [287]. A computer-simulated person within a training system offers a practical solution. It role-plays a situation for learners to learn from, whereby they interact with an interactive computer agent representing a human in a specific training scenario (Figure 2.1). These systems allow learners to practice interactions in a safe, cost-effective, and readily available environment [143, 247, 268].

2.1.2 ARCHITECTURES FOR SOCIAL SKILLS TRAINING SYSTEMS

Rather than building a training system from scratch, it would be efficient for developers and researchers to leverage an architecture that describes the connection and flow between the different components. Integrating different artificial intelligence (AI) technologies and domain knowledge is crucial to the success of training environments [198], making such a system complex by its nature. Thus, an architecture can help designers abstract and understand such complexity. Several architectures can be adapted for training systems, primarily: (1) tutoring models (e.g., [286]), which usually focus on the educational process rather than simulations; (2) virtual human architectures (e.g., [223]); and (3) the integration



Figure 2.1: In a typical agent-based social skill training system, the learner (on the left) engages with a virtual agent role-playing as a human, like a virtual patient, to practice a skill. Based on this interaction, the learner receives feedback, as shown on the right.

of both (e.g., [91]). Other architectures have been proposed in adjacent domains, like in social simulation [54]. However, these simulations aim to study scenarios involving interactions among many agents that mimic human behaviour, such as simulating the behaviour of many people in an evacuation scenario. This makes them different from social skills training simulations, which focus on simulating social interactions to train human learners.

At its core, training systems are systems to teach (the goal) [14] by having a human interact with a simulated environment (the means). Adding tutoring to such simulations could contribute to a guided learning experience, which is more effective than unguided learning [85, 204, 224, 366, 368]. Furthermore, computer-based tutoring can be time- and cost-effective compared to in-house tutoring, e.g., debriefing session [5, 282]. However, we noticed that many existing architectures and systems tend to emphasise either the goal, with less utilisation of the simulation potential, or the means, with less emphasis on pedagogical elements. Yet, these elements combine well and could enhance training system benefits in general. Therefore, we argue for an architecture that supports both the goal and the means of training systems.

Furthermore, we need an architecture to describe and develop systems that capture more complex training simulations. The complexity of these systems spans from modelling passive interactions to simulating human cognition. Examples of the more “complex” systems include negotiation training (e.g., [188]) and job interviews (e.g., [15]). Other systems in this spectrum aim at training learners to understand and change an individual’s cognition or mental state through conversations. These systems allow a human learner to practice how to change a person’s beliefs, attitudes, or emotions on a simulated human to equip the learner with skills to apply in real-life interactions with clients. For instance, simulations might focus on de-escalating aggressive scenarios (e.g., [51]) or persuading a virtual patient to seek assistance (e.g., [99]). These systems specifically train on altering the simulated human’s states, setting them apart from other social skills training. We argue that such systems are among the most intricate to design because developers should replicate components that mimic human thought processes, e.g., modelling changes in

human mental states based on a perceived learner's states. An architecture can provide clearer insights into these systems and their components.

Several surveys have explored training systems, including for virtual patients [30] and those utilizing virtual reality [406]. The closest to our work is by Bosman et al. [48], where they explored 12 agent-based training systems, including those with changing states, and analysed them across six dimensions (e.g., internal reasoning and feedback). Yet, they did not address potential research directions or possible advancements in this area.

2

2.1.3 MAIN CONTRIBUTIONS

In this article, we present ARTES, i.e., the Agent-based training architecture for social skills. ARTES outlines key components that could ideally be present in a system that simulates a training situation with one agent. Being a conceptual architecture [86], it describes the components and the connection and flows between them, regardless of implementation decisions. Thus, the architecture provides a valuable reference point for researchers and developers looking to build upon implementations, while not providing technical implementation information as can be expected in an implementation architecture [189]. One of the key features of ARTES is that it integrates components from two perspectives to maximise learning gains, namely from agent simulations and tutoring systems. In addition, the architecture enables us to identify the important components of training systems, which we can examine further by looking at existing training systems, particularly those focused on changing mental states. By examining the current training systems; offering suggestions for advancements from adjacent fields; linking to learning theories; and discussing future outlooks and research directions, we hope to advance the state-of-the-art in this area.

2.1.4 APPROACH

The article follows an approach similar to Harms et al. [152] and Brabra et al. [55], where they presented a model and discussed current states and possibilities of applying segments of that model from different dimensions. Similarly, we present an architecture and focus on the essential components of training systems. Although our objective was not to conduct a systematic literature review, we defined the architecture and the characteristics by examining a wide range of literature on agent-based training systems and conversational agents in general. To extract recent examples of such systems, we used different sources: (1) by searching for keywords, such as: "training system," "social skills," "agent," and "virtual agent," on Google Scholar and Google, (2) by looking at systems mentioned in three literature reviews on social skills training that fits within our scope [30, 48, 102], (3) by extracting relevant papers published in the Interactive Virtual Agents conference (IVA) between 2019-2022, a period chosen to include also the latest research published after these three literature reviews mentioned, and (4) by looking at reverse citations, where literature references were screened when deemed relevant. The included systems also had to fit within the scope of our architecture, i.e., training individuals in social skills through communicating with an agent. Relevant architectures also inspired ARTES: (a) Architectures of training systems (e.g., Anderson et al. [15], Bosse et al. [50], Kenny et al. [198]); (b) Virtual agents architectures, mainly for input/output components [155, 254], agents' thinking-acting cycles [198] and agents' cognition [302]; (c) Intelligent Tutoring Systems (ITS) models [77, 286, 342]; and (d) An integration of ITS and virtual humans [91].

This paper is organised as follows: We first introduce ARTES, an architecture for agent-based training systems for social skills. We then describe the architecture and its components and present the characteristics by which training systems model the interactions between the learner and the interactive agent, a term which we will use interchangeably with virtual agents in this paper. After that, we identify and discuss ARTES's essential components in training systems, followed by showing their role in four learning theories. Finally, we end the paper by mapping ARTES to eight training systems architectures, classifying the examined systems based on the essential components, linking ARTES to development tools, and discussing further research directions in this area. This research contributes by (a) proposing an architecture for agent-based training systems for social skills, which designers and researchers could use to build on their systems; (b) describing current advances in the aforementioned systems and providing suggestions for what could be integrated from other adjacent domains (e.g., conversational agents); and (c) suggesting future research opportunities for agent-based training systems.

2.2 ARTES ARCHITECTURE AND CHARACTERISTICS OF SOCIAL SKILLS TRAINING SYSTEMS

What should an architecture for agent-based social skills training systems look like? What are the components? And how is interaction with learners characterised? These are the key questions to address now. To explain the abstract concepts in the architecture, we will use the hypothetical example of Lila, a new volunteer in a child helpline. As part of her training, Lila interacts with a simulation system, which includes: (1) a virtual agent simulating a bullied child contacting a children's helpline; and (2) an educational model to manage the learning process. This example is relevant to existing social skills training systems, e.g., to train children in dealing with bullying [25], or to train crisis helpline counsellors [99]. However, it differs in its focus on training a child helpline counsellor to interact with a bullied child.

2.2.1 AGENT-BASED TRAINING ARCHITECTURE FOR SOCIAL SKILLS (ARTES)

Figure 2.2 illustrates the high-level architecture of the virtual training system and how a learner, like Lila, interacts with it. These interactions take place in a communication setting,

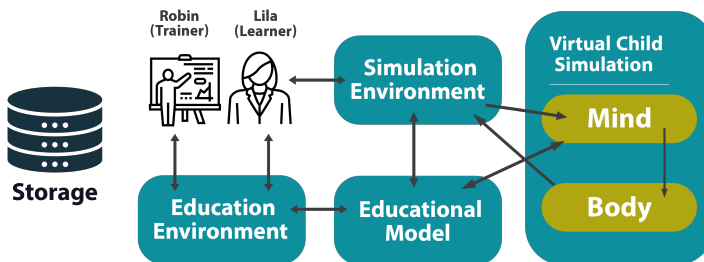


Figure 2.2: A simplified model of the proposed architecture, using the virtual child example.

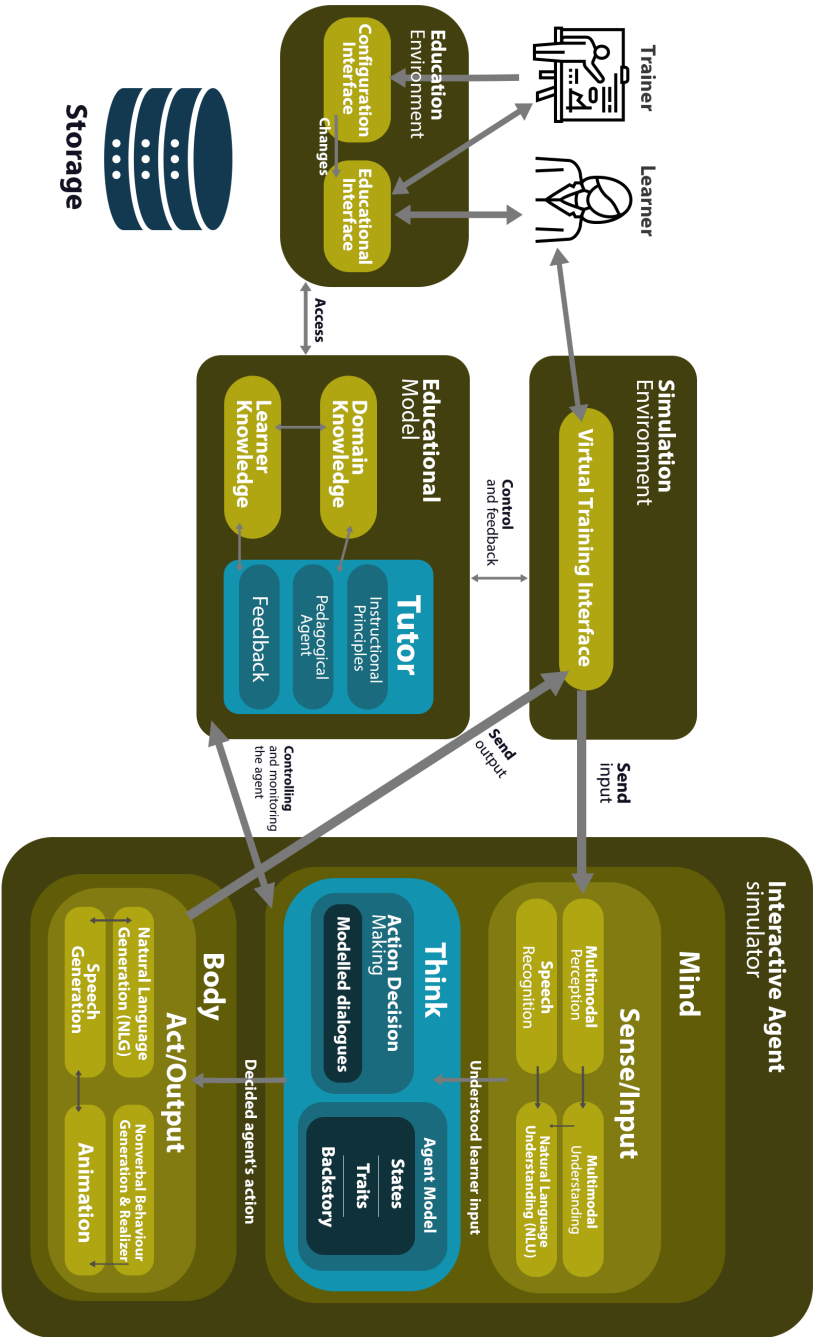


Figure 2.3: General architecture of agent-based training for social skills (ARTES). The components highlighted in blue, “Think” and “Tutor”, are crucial for training learners in social skills. The two components are discussed in detail in this article.

i.e., a scenario. In Lila's scenario, she interacts with a child who has been bullied and wants the helpline to call their school. This simulated scenario allows Lila to practice applying the helpline's communication protocols, which are the steps by which she can help the child explore viable solutions, e.g., to inform a teacher about the bullying. Therefore, Lila practices facilitating a change in how the child perceives and thinks about the situation. Two essential components are needed to facilitate a change: a simulation of the virtual child, which Lila practices with; and an educational model to guide and teach Lila about communication protocols. For the virtual child simulation, Lila interacts with a user interface that resembles the setting used by child helpline volunteers. This includes a chat interface for exchanging short text messages with the child, access to manuals, and a display showing the number of children waiting in the queue. Behind the interface is the virtual child simulation, represented by a mind and a body. The simulated child's mind aims to sense and understand Lila's input and thinks of a reply based on many factors, such as the conversation's history, Lila's empathy, and its belief in the helpline's ability to help. The virtual child uses its body, the interaction modality, to communicate responses. This could have various representation types, such as a physical robot, a virtual character, or, in the case of the virtual child, generating a response and sending it to Lila via the chat interface. Besides a single modal sense-to-act cycle, multi-modal implementations are also possible, e.g., monitoring Lila's emotions expressed via her facial expressions and text input, and also responding across multi-modalities, e.g., a virtual child showing its emotion via body gestures and text responses.

During and after Lila's interactions with the virtual child, she might not know what she did right or wrong. To address this issue, the training tool incorporates an educational model that aims to improve Lila's understanding of the child's behaviour, manage the learning process, and guide her in achieving the learning goals. The educational model uses domain knowledge, such as the helpline's communication protocols, Lila's competence level, and the tutor model, to decide what and how to teach Lila. The educational model could assess Lila's progress, adapt the training scenario or child's personality, and provide Lila with guiding feedback. Lila can access the latter through the educational interface. This interface is also used by Robin, Lila's human trainer, to monitor Lila's progress. Robin can also use the configuration interface to influence the training by, e.g., changing the personality of the virtual child.

The proposed full ARTES architecture (Figure 2.3) aims to present the ideal set of system components that simulate a training situation with one agent, e.g., a virtual child. Lila's example demonstrates that this modular architecture can be tailored to match the simulated training and learning goals. Table 2.1 lists and explains ARTES's main components.

2.2.2 CHARACTERISTICS OF LEARNER-AGENT INTERACTIONS IN TRAINING SYSTEMS

To better understand the interactions between a human learner and an interactive agent in training systems, we identify several key characteristics defining such interactions. Table 2.2 shows them and aligns them with previous works on taxonomies for multi-modality inputs and outputs [117, 369], dialogue management in conversational agents [152, 255] and the existing literature reviews on training systems mentioned earlier. These characteristics and their aspects help explain the components of ARTES and understand the differences

between various system implementations.

Table 2.1: The main components of the architecture.

Component	Description
Virtual training interface	The simulation environment interface, in which a learner interacts with an agent.
Sense/Input	As part of the agent's mind, its main purpose is to sense the user inputs and interprets it. The input could include talking (speech recognition), typing (natural language understanding), happy or sad gestures (multi-modal), or a combination of them.
Think	This part of the mind decides on the agent's next action, once the agent has interpreted the learner's input. For this, it might consider modelled dialogues, agent model (e.g., they have a short temper or negative beliefs of the situation), and how the learner is progressing in training (e.g., the agent takes a less cooperative stance to challenge more experienced learners).
Act/Output	Part of the agent's body. After the agent decides on an action, it "acts" by, e.g., showing sadness (animation) or raising voice (speech generation), which is shown in the training interface.
Domain Knowledge	Contains information concerning the training subject and learning goals. It includes facts, procedures and rules of the topic, such as how to follow a certain communication protocol.
Learner Knowledge and Skill	The learner's current knowledge and skill state in the context of the full domain knowledge-skill set, e.g., to what extent the learner is able to apply a certain protocol.
Tutor	The tutor manages the system's teaching process by considering domain knowledge and the learner knowledge and skill components. The tutor decides on what knowledge or skill to focus on, how to do this, and provides feedback on the learner's performance.
Educational interface	The learner and trainer interact with the educational interface to check the learner's progress and interaction.
Configuration interface	This interface allows a trainer to control the simulation or the educational model, e.g., by changing the agent's mind configuration or tasks.
Storage	The storage keeps data related to the system, and its uses, such as the chatlog or learner's performance data.

Table 2.2: Characteristics of the interactions between a learner and a simulated agent.

Characteristic	Aspect	Definition
Input mode	Verbal	Written learner input.
	Visual	Any visual input such as facial expressions and gaze.
	Sound	Sound input (e.g., speech).
	Haptic	E.g., pressure, selection.
	Others	Sensors, emotion recognition, motion capture.
Verbal interactions input	Closed	Options for the learner to choose from as a reply.
	Open	Free text speech input.
	Both	Have both types of input.
Reasoning	Rule-based	The rules of the virtual agent interactions are predefined.
	Data-driven	The systems learn the rules from conversations (e.g., a dataset).
	Hybrid	Combination of rule-based and data-driven.
Internal state	Defined model	The agent has defined states on how they affect the interactions.
	Limited	The agent has one or two parameters (e.g., trust level) that change.
	None	No explicit representation of internal states.
Agent Knowledge	No history	No memory of past experience is registered.
	Background only	The agent has a history of its backstory.
	Dialogue only	The agent has a memory of past interactions.
	Dialogue and-background.	The agent has a memory of past interactions and their background.
Trainer involvement	Not involved	No human is involved and the agent interactions are automated.
	Configuration	The trainer monitors interaction and configures educational settings.
	Human-initiative	Automated interactions, but the trainer can intervene and control.
	Agent-initiative	The agent asks the trainer to take control (e.g., making a decision).
	Wizard of Oz	A human understands the input and controls the output of the agent.
Output mode	Verbal	The agent utterances.
	Visual	Any visual output appearing on the agent (e.g., animations).
	Sound	Sounds the agent outputs (speech or non-speech).
	Others	Outputs that can not be seen or heard (e.g., haptics).

While any combination of these characteristics can theoretically occur, certain patterns are more likely based on factors such as the agent's roles, technology used, and training scope. When considering the case of Lila's interaction with the virtual child, the interaction takes place through open written inputs, where Lila writes in a conversation text field. The child's response is provided through verbal textual output on the interface, and the reasoning is based on a hybrid approach between rule-based and data-driven, allowing for training in a safe, explainable, and controlled environment. The child also has defined internal states and beliefs that change while talking with Lila, as well as a background story of being bullied at school, and memory of past conversations. The trainer, Robin, can modify the virtual child's configuration by changing its belief values, making the child more

or less adamant. In other cases, the agent simulation may have different characteristics. For example, a visual embodiment of a virtual recruiter for training on non-verbal cues, direct rule-based reasoning for answering questions in history-taking, or an agent-initiated involvement to demonstrate specific cues of a virtual patient with depression.

2

2.3 KEY SOCIAL SKILLS TRAINING COMPONENTS

The ARTES architecture includes components that are also included in non-training systems, such as multi-modality or dialogue systems. As these more generic components have been explained elsewhere (e.g., Harms et al. [152], Turk [369]), we will now focus our attention on the “Think” and “Tutor” components. Both components play a pivotal role in training learners’ social skills to understand and change an individual’s emotions, cognitions, and behaviours consequently. Table 2.3 shows the sub-components of these two components.

Table 2.3: Overview of Think and Tutor sub-components, which we examine further in this article.

Examined Components	Description
Think:	An agent decides on an action based on the understood input.
<i>a) Action Decision Making.</i>	Controls how the training scenario proceeds, i.e., how an agent decides what their next step would be based on the overall interaction structure.
<i>b) Agent model.</i>	Describes the model of the agent with different parameters and states. This could include its emotions and cognition about itself, the learner, and the world. This, in turn, influence the agent’s future behaviour.
Tutor:	What educational information to present and how.
<i>a) Instructional principles.</i>	A training system includes principles to teach knowledge, which could affect the agent’s interactions with the system. It is related to learning techniques and the learner’s state.
<i>b) Feedback.</i>	Generates and displays feedback to reflect on the learner’s performance, e.g., being empathic, or to explain the agent’s behaviour, e.g., the changes in agent beliefs.
<i>c) Pedagogical agent.</i>	A virtual trainer agent guiding or reflecting upon the learner’s interaction with the simulated agent. It can have basic or complex interactions, e.g., with the “thinking” elements.

2.3.1 THINK COMPONENTS

Most training systems we studied were limited in their simulation of the agent's thinking. We believe there is a potential for improvement, as by showing the change in the agent's belief states, learners can construct a mental model of the agent's thinking process, from which they can learn the consequences of their actions. In other words, they can construct a mental model of the situation they operate in [190]. For Lila, this means she could understand how the child's emotion and thinking changes when she follows the helpline protocol. In this section, we will discuss the elements that affect the agent's action decision, reflecting on the proposed architecture.

SOURCES OF INPUT

Looking back at the model in Figure 2.3, we distinguished three information sources upon which an agent bases its actions: what Lila has said (information provided by sense/input component); what the child is thinking about itself or the world (the agent model); and what is an appropriate learning experience for Lila (input from the educational model). Let us start with the first one. The agent considers Lila's input based on what the two have said so far, e.g., did the child pose a question, and is this Lila's answer? If the child is happy and trusts Lila, the agent might consider her answer instead of rejecting it. In other words, the agent's decision depends on the agent's cognitive and emotional states, i.e., the second source of input. So, if Lila showed empathy, the child's trust in Lila would increase; thus, the agent would be more willing to open up. Finally, Lila might be cognitively and emotionally ready for a more challenging twist in the scenario. Sensors indicate she is relaxed, and the learner knowledge and skill information indicate she has sufficient understanding for the next step in her training. Therefore, the educational model might change the parameter setting of the Action Decision Making component on the fly, for example, making the child more susceptible to any suggestion of rejection by Lila. This third source, therefore, influences the agent's decision-making on a meta-level, by changing the rules it applies.

BASIC DECISION MAKING

With the three input sources, an agent's thinking could vary widely, from being simple options-based, matching inputs to direct outputs, to having a complicated structure of decision-making that includes dialogues, goals, and emotions that adapt to the learner's state. A direct thinking process might be suitable for training systems where a change in an agent's beliefs or emotions as part of the conversation is less relevant to the agent's decision-making. For example, to train student doctors in asking specific appropriate questions and collecting information for diagnosing [338]. However, such dialogues are experienced as being static and inflexible [48]. Therefore, some proposals have been made to tackle this. For example, Hirumi et al. [167] structured the questions that could be asked by learners into topics, e.g., about the patient's medical or social history. Dialogue managers [152] go beyond option-based solutions, as they can consider the dialogue history and state when deciding on an action. Take, for example, the following interaction: “[Child] I'm bullied at school, can I talk about that?; [Lila] Sure, I am here to listen; [Child] Thanks. Do children often call about bullying?; [Lila] yes, you are not the only one”. Here, the dialogue manager is running a script and, based on Lila's input, gives an appropriate response and moves on. Another extension is to add one or two state variables about agents, e.g., a trust

meter [332]. It is an initial attempt to have learners consider the consequences of their actions.

COGNITIVE MODELS

More complex agent models can be used to elaborate the thinking, e.g., by simulating a virtual child's mental model that reaches decisions based on interactions with Lila. These models are often referred to as cognitive models that simulate human decision-making processes [362]. These cognitive models can be application-specific or can be built upon general theories. Application-specific cognitive models, such as a virtual suspect for interrogations that bases its answers on interpersonal relations [65], have the advantage of addressing unique simulation aspects, e.g., states which might not be included in general theories. However, they require more work from domain experts to simulate thinking, agent states, and their interconnections. In the case of the virtual child, this involves identifying relevant thinking processes for training, potential cognitive and emotional states for the child, and how these two can be linked together. On the other hand, general cognitive models describe how the agent's state varies, allowing developers to build upon a predefined flow. One well-known and utilized cognitive theory in conversational agents is the belief-desire-intention (BDI) model. This model is built on three aspects: (1) the agent's beliefs about themselves, the learner (i.e., Theory of Mind [226]), and the world; (2) the agent's desires regarding their goals; and (3) their intentions as the actions they wanted to execute [136]. For instance, when Lila shows empathy towards the virtual child, the child's belief in Lila's understanding of their situation increases, making the child more open. Employing BDI in such a model could simulate human decision-making, adapt to changes in the environment, and directly explain the child's decisions and changes in internal states. In their survey of BDI agents for social simulation, Adam and Gaudou [2] argued that BDI models are well-suited for agents in training systems. However, they also discussed drawbacks, with the most relevant for training systems being the need for expert input to model agents' values. BDI has been utilized in several training agents (e.g., McShane et al. [253], Muller et al. [272], Oijen et al. [289]), and proposed for others (e.g., Baptista et al. [28], Van der Zwaan et al. [379]). The model's flexibility allows for extensions to include social concepts in the agent's mind, such as agent trust [2], which would be valuable to train learners on specific social skills and how such concepts affect the agent's decision-making.

COGNITIVE ARCHITECTURES

Cognitive architectures offer another approach to modelling human cognition in training settings. Although some cognitive architectures are based on BDI, researchers differentiated BDI from cognitive architectures in that the former is based on a philosophical take on behaviour, while the latter considers describing lower-level cognitive processes to simulate a human mind [2]. Therefore, we will consider them both as methods of cognition modelling in the agent's thinking. Cognitive architectures aim to achieve general human intelligence capable of reasoning, insight, behaviour adjustment, and self-reflection [212]. Architectures vary in decisions about actions, typically based on factors from their internal memory and interactions, such as physiological needs, utility, state, and relevance. However, we found limited social skills training applications based on surveyed cognitive architectures [212, 405]. One exception is a 2003 implementation of Soar [216] to train army personnel to

make decisions in a critical social situation [164]. The limited examples might be because cognitive architectures aim to simulate general human intelligence and not to “be an actor” with emotions and states in a specific role-play training scenario. Moreover, they are not as easy to implement as BDI [2]. While cognitive architectures may not be the primary choice for training systems with changing agent states, they might have a potential added value when continuous adaptability or self-reflection is needed, depending on the agent’s past experiences. For example, when building a general virtual human for helpline counsellor training that can be utilised in children’s helplines.

MODELLING EMOTIONS

So far, we have only discussed the cognitive part of the agent’s mind. An additional layer to the agent model in a training system is modelling the agent’s emotions, often referred to as affective architectures [302]. These architectures extend cognitive models, such as BDI, to include more elaborated states or traits to the agent, e.g., by adding emotions, personality, attitude, and mood [290]. As a result, the agent’s emotions can influence their decisions and beliefs, leading to varied reactions based on their feelings. Incorporating such feelings in the virtual child example can help Lila understand the child’s internal states and emotions and how the interactions affect them, e.g., when the child might feel frustrated and how to address it. Emotions also enrich the simulation, particularly when the learning outcome involves understanding or changing the agent’s emotions. One example is TARDIS [15], which has an affective architecture. In this job interview training system, the virtual recruiter’s emotions impact its decision-making, e.g., becoming more aggressive as they feel angry. Numerous affective models, primarily based on BDI [290], integrate cognitive modelling with emotions to build a training agent. For instance, EMA [245] was employed in a stressful decision-making scenario [359], and FATiMA [101] to build a job interview scenario [248].

DATA-DRIVEN APPROACHES

Agents in training systems usually have rule-based reasoning. Alternatively, data-driven approaches, i.e., a model trained on corpora of conversations [152], may also prove helpful. There are many challenges, though, to having data-driven training systems, including the scarcity of conversational data tailored to the simulated scenario, concerns about data accuracy, and technical limitations such as limited understandability of human emotions. For example, building a data-driven virtual child would raise questions regarding sensitivity, availability, and privacy when collecting conversational data from children and helpline volunteers. Human-in-the-loop machine learning could offer a solution when a human trainer acts as the agent and responds to learners’ interactions to create a corpus. In Lila’s case, this would involve Robin controlling the virtual child. This method, akin to what was used in a storytelling chatbot [181], could help train the virtual child on actual data for different background stories and interactions. However, it requires Robin’s commitment and manual effort to define the child’s states in order to develop a human interpretable agent model. Reinforcement learning offers another alternative, where the model of the agent is trained based on user interactions, and the agent’s correct behaviour is rewarded [395]. While reinforcement learning has been utilized in related domains, such as healthcare assistants [403], implementing this approach in a training system like Lila’s may pose challenges. These challenges can arise when an agent needs to have

consistent and predictable agent behaviour across different learners. Furthermore, defining suitable reward functions for simulation in such a sensitive scenario, which requires interdisciplinary expertise, poses its own hurdles. Therefore, human experts could check such rewards to determine if the virtual child exhibits appropriate behaviour during the communication protocol phases or to assess whether Lila gained the correct learning skill, such as showing empathy to the child.

2

TRAINER'S INVOLVEMENT

Human trainers could also influence the agent's decision-making process. ARTES captures this by incorporating the interaction flow from the trainer, in our case, Robin, to the educational model and then to the agent's thinking. Aside from the Wizard of Oz approach, where a human trainer substitutes the entire agent cognition, Table 2.2 illustrates two ways in which a human trainer can impact the agent's action decision: based on human or agent initiative. Human initiative occurs when the trainer acts as a supervisor guiding the agent and intervening when necessary, for example, if the virtual child does not exhibit the expected behaviour or if Robin wishes to control parts of the agent (e.g., Hartanto et al. [153]). On the other hand, agent initiative has the agent asking Robin to choose on its behalf, substituting a part of the agent's cognition with the trainer's input. The latter is used when the agent is stuck in a conversation or faces a moral dilemma [377]. While this approach requires a human trainer on standby during the training session, it can also be utilized in a data-driven approach to training the agent model, where the agent learns from the trainer's decisions by requesting their input.

2.3.2 TUTOR COMPONENTS

The tutor model is an intermediate between domain knowledge and learner knowledge in the educational model. The tutor manages the learning experience by supervising learning tasks, providing feedback, and maintaining learner engagement and motivation [53]. Unlike agents with their human-like manifestation, tutors in intelligent tutoring systems could manifest themselves in many forms. These can be colour-coded feedback indicators, such as red and green symbols showing the learner's performance or topic recommendations. Many models have been reported for tutors, detailing their logic and presentation style for giving feedback or guidance to a learner [13]. Having such a model in a training system is helpful to educate learners, as the lack of guidance could leave the learner only with the simulation experience, limiting its effectiveness [158]. Consider, for instance, if Lila directly instructs the child to ask the school for help. Such an approach contradicts the helpline's guidelines [96], where a counsellor is expected to coach the child to come up with that solution rather than telling the child directly what to do. A tutor could intervene in such cases, showing Lila how her approach deviated from the helpline's guidelines.

Adding a tutoring model could enhance the training impact by giving the right feedback at the right moment, thereby integrating guidance naturally into the training system's teaching approach. In short, this model is essential for offering a guided learning experience. Designers could look at a number of learning theories [113] established by different schools of thought to utilise them in training systems, thus, facilitating the learning process. On the other hand, training systems are typically designed to teach a specific set of skills.

Therefore, linking the learning theories and training systems components clarifies the latter's roles, allowing for a more targeted implementation. But before exploring this link, we must first understand the tutor's subcomponents.

THE THREE SUBCOMPONENTS

The tutor model (Figure 2.3) is responsible for teaching the learner, and has three components: instructional principles, a pedagogical agent, and feedback. The first component, instructional principles, or as some refer to it, instructional strategies [393], entails psychological techniques the training system can employ to deliver the training. Take, for example, the scaffolding strategy [376]. Here, the system initially offers Lila extensive help, then gradually reduces support as she gains experience. Next is the pedagogical agent, who behaves as a social actor as it simulates a teacher guiding the learner, e.g., as a person offering guidance and assistance to Lila. Although pedagogical agents are perceived positively in training systems (e.g., Murali et al. [273, 274], O'Brien et al. [296]), they are often not included. For example, Table 2.6 shows that in 43 systems we studied, only 12 had a pedagogical agent. The third tutoring component is feedback, which can give Lila insight into her performance. The feedback can be based on the modelled knowledge, e.g., to tell Lila what she did wrong and offer hints, or the agent's states, e.g., why the virtual child does not trust Lila. Yet, most examined systems primarily focus on the former (Table 2.6).

The tutor model can interact with (a) the agent simulation, e.g., for monitoring belief changes in the simulated child that needs explaining, (b) the domain and the learner knowledge, e.g., for determining when Lila's training should move on to a more challenging situation, and (c) the human teacher, e.g., Robin configuring the tutor. Similar to ITSs [380], the tutoring model operates within two learning cycles: the within-session cycle, e.g., for providing hints during a session, and the between-session cycle, e.g., for deciding on new scenarios.

LINKING TO LEARNING THEORIES

Tutor components can represent and facilitate the implementation of learning theories, thus supporting their incorporation and abstraction within ARTES. Four prominent learning theories exist today that explain how skills are acquired: behaviourism, cognitivism, social cognitive theory, and constructivism [113, 334]. Table 2.4 highlights the roles of four software components within the context of these learning theories. They are the three tutor components and the virtual agent simulation. The first software component, instructional principles, derives its role directly from the core ideas of learning theory, subsequently guiding the design of agent simulation. Take, for example, behaviourism. The instructional principles focus on forming appropriate stimuli-response relations; thus, the simulation's role is to offer realistic stimuli. This flow similarly applies to the roles of the pedagogical agent and feedback, which, in our example, work to reinforce desirable responses to stimuli through guidance and feedback. Now, let us look at the learning theories in more detail.

Behaviourism can be applied to teach facts and tasks through repetition and memorization. This approach supports reinforcement and guidance in accordance with instructional principles. Therefore, the realism of the agent's actions and consequences are important as they establish stimuli for which a response needs to be trained. To strengthen this

Table 2.4: The roles of ARTES components within four primary learning theories. These components comprise the three tutor elements (instructional principles, pedagogical agent, and feedback) and the agent simulation. The roles explain how these components fit within each learning theory.

Learning Theory	Instructional Principles	Agent Simulation	Pedagogical Agent	Feedback
Behaviourism: Reinforcement to shape desirable behaviours	Use of techniques to guide learners behaviour	Realism of stimulus	An entity that provides feedback or assistance	Provide reinforcement feedback
Cognitivism: Focus on mental process to acquire knowledge	Use of techniques to guide learners knowledge development	Simulating several situations to learn from	Teaching and explaining	Explaining situations and learner's decisions
Social Cognitive Theory: Learning through social observations	Offer suitable examples of interactions	Simulating several situations to learn from	Interacting with the agent simulation	Explaining situations
Constructivism: A learner constructs their own understanding and knowledge	Facilitate the learning environment	Explorative situation	Mentoring	Used alongside instructional principles to guide learners

stimulus-response, the pedagogical agent and feedback components should reinforce positive or negative behaviours. Some examples of adopting behaviourism are showing angry responses and allowing scenario repetition following incorrect actions by the learner [146], as well as a virtual coach giving a thumbs-up for correct actions [273].

Cognitivism focuses on constructing a learner's cognitive process and knowledge, which requires an understanding of the relevance and application of targeted skills. Implementing instructional principles, such as information chunking, can build this understanding. Accordingly, an agent simulation with several learning situations is needed to demonstrate the use of a targeted skill, which is further explained by a pedagogical agent. Feedback can then be provided based on the learner's interactions with the simulation. Examples of cognitivism in training systems include: adapting a scenario difficulty to the learner's current knowledge [51], a pedagogical agent that explains the next task [274], or receiving reflection about the learner's performance [15].

The social cognitive theory emphasises learning through observation of peers or coaches. In this perspective, the instructional principles could demonstrate successful interaction with a virtual agent to train the learner. Therefore, the examples should cover different learning situations. Several systems have implemented this approach, including a pedagogical agent that provides examples of taught skills [274], or a system that allows learners to view video recordings of mastered skills [363].

As for *constructivism*, learners should be able to construct their knowledge, e.g., through self-reflection [211] or discovery learning [250]. In training systems, this translates to exploring virtual agents' scenarios to create their interpretations and communication styles

and learning from mistakes. This approach requires a diverse and rich set of scenarios in the simulation. At the same time, the instructional principles can encourage exploration by facilitating and personalising the environment presented to learners. Examples of applying this theory include providing a “practice” mode to support scenario exploration and learning [15], a virtual coach providing insights into the virtual patient’s thinking [296], and offering guiding feedback by highlighting important discoveries during the interaction with the agent [206].

ALTERNATIVE TUTOR STRUCTURE THROUGH DATA-DRIVEN APPROACHES

Utilising data-driven techniques in the educational model makes it harder to distinguish between the different educational model components. One example is the Korbit learning platform [208, 353]. Korbit aims to provide instruction in various subjects and skills, employing a conversational interface that utilizes reinforcement learning and machine learning to personalize the AI tutor’s pedagogical intervention. It utilises machine learning to generate hints based on learner questions. It also retrieves explanations from sources like Wikipedia. However, this technique requires extensive data sources that are typically only available for general educational topics, making it less suitable for training systems that focus on specific domain aspects and train learners accordingly. Nonetheless, Korbit’s approach can be valuable, allowing for adapting instructional principles based on learner profiles and previous interactions with the system. This requires a more elaborate training system with diverse scenarios or tasks to build the learner’s profile and history. Among the examined training systems, only one system integrated reinforcement learning into its tutoring approach [138], where it manages feedback and prompts self-reflection on incorrect user inputs. This implementation showcases a merging between different components from our model, specifically the instructional principles, feedback, pedagogical agent, and learner knowledge.

2.4 RELATION WITH THE EXAMINED SYSTEMS, OTHER ARCHITECTURES AND TOOLS

Having described ARTES, it is time to see how well the architecture captures and resonates with existing systems and architectures, supporting that ARTES is versatile and agnostic. This assessment provides insights into its deployment and adaptability across different situations. Specifically, whether ARTES shows relevance and generalisation and whether its components match those of other architectures. Moreover, looking at how the components were implemented, e.g., if the input type was choice-based or open-ended, gives an insight into how concepts discussed earlier are actually implemented in systems. Next, we will examine how well current software tools support ARTES-based system development. All these points are, therefore, systematically examined here.

2.4.1 MAPPING ARTES TO SOCIAL SKILLS TRAINING ARCHITECTURES

How complete is ARTES in comparison to existing architectures? To answer this question, we evaluated the coverage of ARTES against eight architectures suitable for training systems. We did this by mapping the components of the eight architectures to those of ARTES (Figure 2.4). The mapping offers insights into components commonly defined in

training or tutoring systems, compares ARTES's components with other architectures, and highlights any component that ARTES might have overlooked. The eight architectures we analysed are all conceptual architectures deemed relevant to ARTES and applicable within training systems. These architectures are (a) two implementations of training systems (aggression de-escalation [51], and TARDIS [15]), (b) three virtual agent simulations (virtual human integrated architecture [198], the virtual human architecture [223], and FATiMA Toolkit architecture [248]), (c) two tutoring systems (the general ITS architecture [342], and a pedagogical agent in an ITS system [77]), and (d) an architecture integrating ITS with virtual agents [91]. Although the first two architectures were developed to describe specific systems, we think their information is general enough to be applied to training systems. Still, the mapping accuracy is limited by the clarity to which the architectures were described in the literature and potential individual interpretation bias. To mitigate the latter, two coders with a computer science background mapped the eight architectures to ARTES independently, which resulted in a substantial agreement (Cohen's $\kappa = 0.7$) [217]. The coders then discussed disagreements to reach a consensus.¹

We could directly map all ARTES components and layers to at least one component from the eight architectures, indicating that ARTES components are relevant to training systems. This gives more credibility to ARTES's components. On the other hand, we did encounter six components from the mapped architectures that did not have suitable matches in ARTES. Two reasons for this are: i) the components were describing the interaction process in a concealed "black box" that generalised across distinctive components and their relations (*system control* that handles interactions between all components [342] and *communication bus* between all components [198]); and ii) that they were from other architectural perspectives than conceptual architecture (a component for the general software *tools* [198] and *real-world environment* that represents the real world [223]).

On average, ARTES matches or exceeds the detail level of the mapped architectures, such as FATiMA, which has more configuration components but fewer Mind and Body components. This also includes specialised architectures like Virtual Humans for agent simulation and ITS architectures for educational systems. Figure 2.5 shows the depth these eight architectures cover, based on the mapping in Figure 2.4. There are three main clusters in the figure: (1) in the top left corner are architectures that mainly focus on the agent's simulation (the means); (2) in the bottom right corner are those focusing on the educational model (the goal); and (3) in the middle, the integration of both in well-defined systems. However, none of these eight architectures addresses both aspects as deeply as ARTES (located in the top right corner), indicating that these architectures often focus more on either one of the dimensions². In practice, training systems like TARDIS and aggression de-escalation often incorporate components from both educational and agent simulation models, supporting our vision to merge tutoring and agent models. Thus, ARTES demonstrates coverage with both specialized and broader system architectures, underscoring its completeness.

¹The analysis of the double-coding is available through the 4TU research data repository: <https://data.4tu.nl/datasets/a037707b-5d70-410a-b6e1-9ee4a694b1d7>

²Note that the mapping does not evaluate the quality or implementations of the architectures, but rather showcases ARTES's coverage. These architectures may support additional functionalities and libraries beyond what is reported in the examined literature.

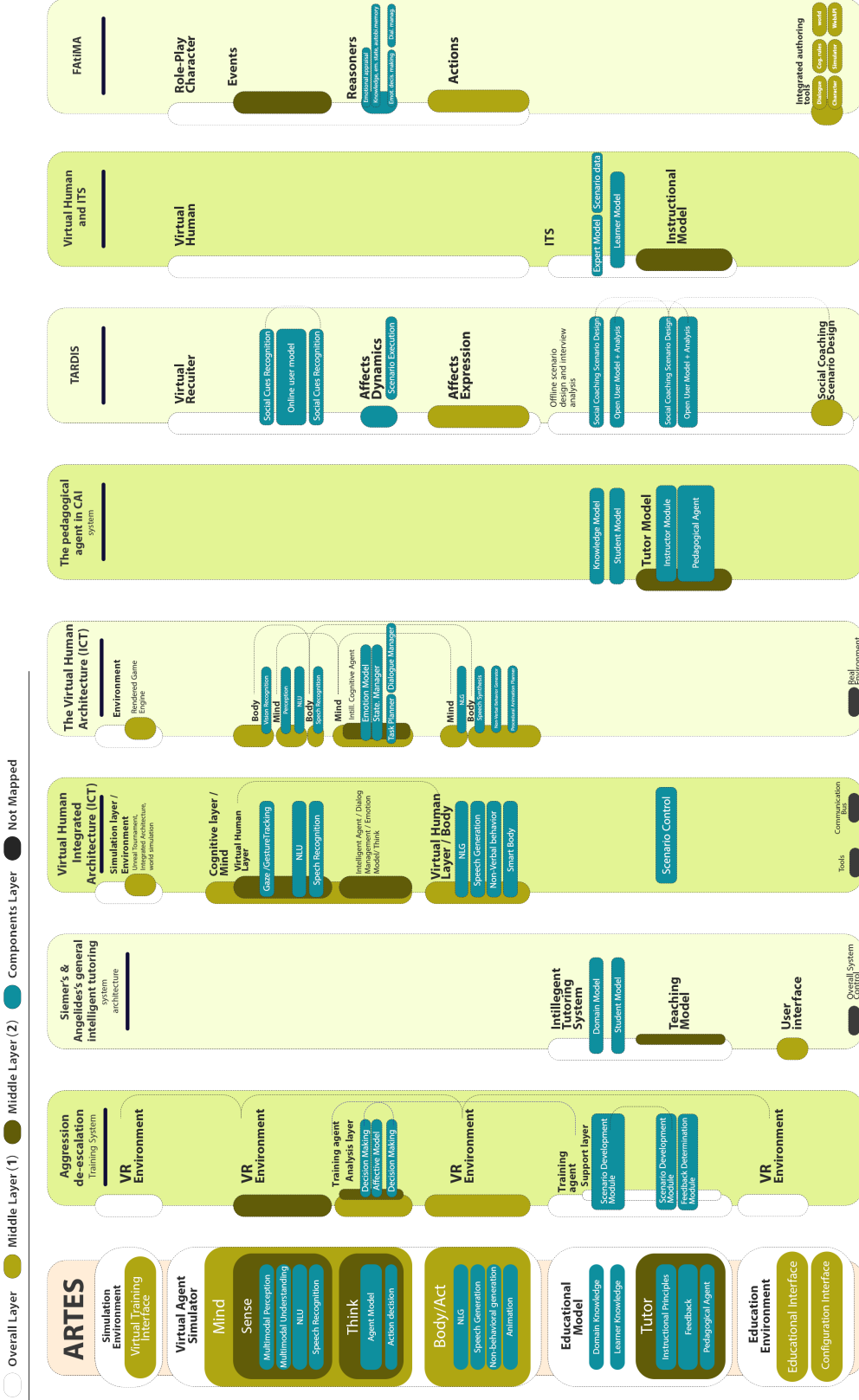


Figure 2.4: Mapping components between ARTEs and eight other architectures. Columns represent different architectures, while rows show ARTEs components (on the left) and their counterparts in other architectures. Dotted lines connect component names that map to multiple ARTEs components. Colours indicate layers and black areas highlight components absent in ARTEs.

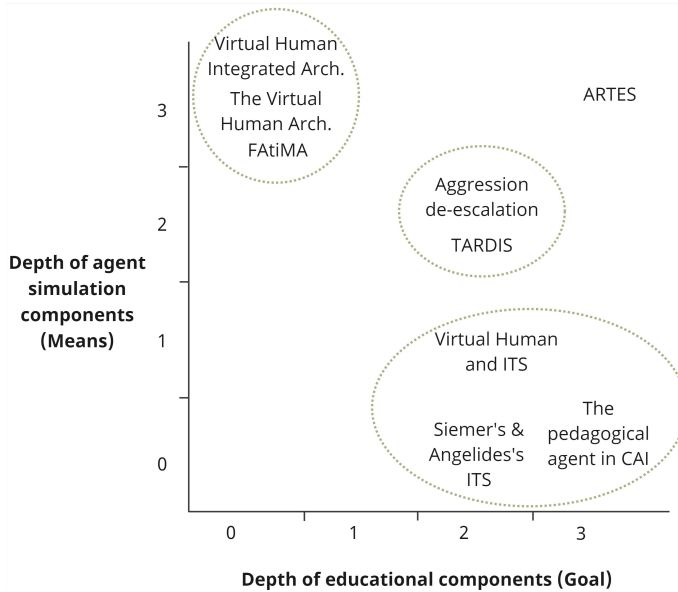


Figure 2.5: Assessing the level of depth for the eight mapped architectures (Figure 2.4). We determined the level of depth by calculating the median depth level of the components of an architecture for the two dimensions: the agent simulation model and the educational model.

2.4.2 EXAMINED SYSTEMS

To further explore the generalisability and relevance of ARTES components, we examined 43 training systems. Specifically, we studied if we could identify the five ARTES important components, as highlighted in section 2.3, within the examined training systems, and if the concepts presented in this article resonate with existing systems. To do that, we formulated Table 2.5, which presents the categorisation criteria for components' properties, such as input type, learning theories, and pedagogy format. We based these categorisations on the ideas we discussed previously in section 2.3. Also, B.1 visually explains in more detail two classifications of the “decided by” property of the Action Decision Making component (Figures B.1a and B.1b) and the “structure” property of the Agent model component (Figures B.1c and B.1d). Then, we applied the categorisations Table 2.5 to the examined systems, which resulted in Table 2.6. Just as with mapping the architecture components, we cannot guarantee the accuracy of our interpretations based solely on the descriptions of these systems. Therefore, we conducted reliability testing to examine this bias. Five secondary coders classified a total of 41 of the 43 systems, with the remaining two systems used to train them in coding. There was moderate inter-rater reliability between the first coder—who assessed all systems—and the secondary coders (average Cohen's $\kappa = 0.55$, range: 0.49–0.63) [217]³. The coders discussed the ratings to reach a consensus.

Table 2.6 maps the five components across the examined systems. For Action Decision

³The analysis of the double-coding is available through the 4TU research data repository: <https://data.4tu.nl/datasets/a037707b-5d70-410a-b6e1-9ee4a694b1d7>

Table 2.5: Categorisations of the different implementations for the examined components, based on the discussion in section 2.3.

Components				
Action Decision Making	Agent model	Feedback	Instructional principles	Pedagogical agents
<p> Decided by:</p> <ul style="list-style-type: none"> (1) The teacher actor. (2) Direct input to output mapping. See Figure B.1a. (3) A scenario sequence is defined. See Figure B.1b. <p> Also affected by:</p> <ul style="list-style-type: none"> (1) Agent model. (2) Educational System. <p> Input type:</p> <ul style="list-style-type: none"> (1) Choice-based. (2) Open-ended. 	<p> Structure:</p> <ul style="list-style-type: none"> (1) Fixed changes in the states. (2) Limited states change. See Figure B.1c. (3) Representation of changes in many states. See Figure B.1d. <p> Represented states:</p> <ul style="list-style-type: none"> (1) Changing cognition. (2) Changing emotions. <p> Model of:</p> <ul style="list-style-type: none"> (1) The agent itself. (2) The learner. (3) The world. 	<p> Moments of feedback:</p> <ul style="list-style-type: none"> (1) During interactions. (2) After interactions. <p> Content about:</p> <ul style="list-style-type: none"> (1) The agent's mind (2) The modelled knowledge <p> Covers:</p> <ul style="list-style-type: none"> (1) One session or task. (2) Multiple sessions or tasks. 	<p> Learning theories:</p> <ul style="list-style-type: none"> (1) Elements of behaviourism. (2) Elements of cognitivism. (3) Elements of social cognitive theory. (4) Elements of constructivism. <p> Extent:</p> <ul style="list-style-type: none"> (1) Revolves around one dimension of principles. (2) Multiple dimensions. 	<p> Representation of a trainer:</p> <ul style="list-style-type: none"> (1) Through text or sound. (2) Embodiment. <p> Pedagogy format:</p> <ul style="list-style-type: none"> (1) Static. (2) Interactive agent. <p> Role:</p> <ul style="list-style-type: none"> (1) Reflecting and feedback. (2) Learning, e.g., describing concepts or answering questions. (3) Assisting. (4) Social actor.

Making, 49% of systems utilized a scenario sequence to guide agent actions, while 42% employed input-output processing. Additionally, 28% of agents incorporated a representation of internal states. Surprisingly, only 70% of the systems integrated feedback components, where 14% provided feedback only during the interaction, 33% only post-interaction, and 23% during both phases. Regarding instructional principles based on learning theories, Behaviourism was most prevalent at 81%, followed by Cognitivism (60%), Constructivism (49%), and Social Cognitive Theory (12%). This prevalence might stem from training systems' aim to train learners on social situation responses, often focusing on behaviour formation. Lastly, a mere 28% of the systems featured a pedagogical agent, either as text (9%) or an embodied agent (19%).

Table 2.6: Categorising the examined training systems based on Table 2.5.

System	Training type	Examined Components				
		Action Decision Making	Agent Model	Feedback	Instructional Principles	Pedagogical Agent
Bosse et al. [50, 51]	Aggression de-escalation	3, 1,2, 1	3, 1,2, 1,2,3	2, 2, 1	1,2,4, 2	-
MPathic-VR [146, 213]	Medical communication	3, 1	1	2, 2, 1	1,2, 1	-
Murali et al. [274]	Counseling	3, 1	1	-	1,2,3, 1	2, 2, 2
TARDIS [15]	Job interviews	3, 1,2, 2	3, 1,2, 1,2,3	2, 2, 1	1,2,4, 2	-
NERVE [166, 167])	Interview and diagnosis	2, 1,2	1	1,2, 2, 1	2,4, 2	-
ASST [363]	Communication skills	2, 2	1	2, 2, 1	2,3,4, 2	-
CuCoMaG [104, 295]	Customer service	3, 2	1	2, 2, 1	1,2, 2	-
Virtual Suspect William [64, 65]	Suspect interrogation	1,2, 1, 2	3, 1,2, 1,2,3	2, 1, 1	1,4, 1	-
Friendly Face [273]	Public speaking	2, 2	-	1, 2, 1	1, 1	2, 2, 1,3,4
Nakash et al. [278]	Suicide prevention	1, 2	-	-	1, 1	-
Communicate! [184]	Communication Skills	3, 1	1	2, 2, 1	1,2, 1	-
Schoenthaler et al. [332], Albright et al. [10]	Medical communication	3, 1, 1	2, 1,2, 1,2	1,2, 2, 1	1,2,3, 2	2, 1, 1
Maicher et al. [239]	History taking	2, 2	1	-	1, 1	-
O'Brien et al. [296]	Suicide prevention	2, 1, 1	2, 1,2, 1	1,2, 1,2, 1	2,4, 2	2, 1, 1,3
Muller et al. [272]	Sales training	3, 1, 1	3, 1, 1,3	1, 1, 1	2,4, 1	-
Yao et al. [400, 401]	Medical empathy training	2, 2	1	2, 2, 1	2,4, 2	1, 1, 3
ELECT BiLAT [218]	Intercultural negotiation	2, 1,2, 1	2, 1,2, 1,2,3	1,2, 1,2, 1	1,2,4, 2	2, 1,2, 1,2,3
Szilas et al. [361]	Caregivers training	3, 1	1	-	1, 1	-
Ochs et al. [287]	Medical communication	1,2, 2	-	2, 2, 1	1,4, 1	-
Georgila et al. [138]	Counseling skills	3, 1	1	1,2, 2, 1	1,2, 2	1, 1, 1,3
Ziebarth et al. [408]	Medical Interviews	2, 1, 2	2, 1,2, 1	1,2, 1,2, 1,2	1,2,4, 2	-
Adewole et al. [4]	Cultural Awareness	3, 2	1	1, 2, 1	2,4, 2	1, 1, 1,3
Bánszki et al. [27]	Medical education	1, 2	-	-	1, 1	-
UT TIME Portal [409]	Medical communication	3, 1	1	1,2, 1,2, 1	1,2,3, 2	-

Continued on next page

System	Training type	Examined Components				
		Action Decision Making	Agent model	Feedback	Instructional Principles	Pedagogical Agent
Zlotos et al. [411]	Medical communication	2, 1	1	-	1, 1	-
AdaCoach [301]	Customer service	3, 2	1	1, 2, 1	1, 1	-
Dupuy et al. [108]	Medical interviews	3, 2	1	1, 2, 1	1, 1	-
Mission Rehearsal Exercise [164]	Decision-making skills	3, 1, 2	3, 1,2, 1,2,3	-	1,4, 1	-
Peddle et al. [299]	Medical communication	3, 1	1	-	1, 1	-
deLearyous [374]	Communication skills	2, 1, 2	3, 1, 2	-	1, 1	-
Washburn et al. [388]	Medical education	2, 2	1	2, 2, 1	1,4, 1	-
Demasi et al. [99]	Hotline counselling	3, 1,2, 2	1	-	1,4, 1	-
INOTS [160]	Interpersonal skills	3, 2	1	2, 2, 1	2,3, 1	-
Virtual-Suspect [39]	Suspect interrogation	2, 1, 2	2, 1, 1	-	1,4, 1	-
Mell and Gratch [256]; Johnson et al. [188]	Negotiations	2, 1,2, 1,2	3, 1,2, 1,3	2, 2, 1	1,2,4, 2	-
Jacklin et al. [179, 180]	Shared decision making	3, 1	1	2, 2, 1	1,2, 1	-
Sveinbjörnsdóttir et al. [357]	Training teachers	2, 1	1	1, 2, 1	2,4, 2	-
DialogueTrainer [100] ⁴	Several skills	3, 1	1	2, 2, 1,2	1,2, 1	2, 2, 2
SIMmersion [343] ⁴	Several skills	2, 1	1	1,2, 1,2, 1	1,2,4, 2	1, 1, 1
Kognito [209] ⁴	Several skills	3, 1	1	1,2, 1,2, 1	1,2,4, 2	2, 1, 1
Mursion [275] ⁴	Several skills	1, 2	-	- ⁵	1,2,4, 2	2, 2, 4
VirtualSpeech [385] ⁴	Several skills	2, 2	1	1,2, 2, 1,2	1,2, 1	-
CleVR [87] ⁴	Several skills	1, 2	-	-	1,4, 1	-

⁴ These systems are commercially available, and were categorised based on information from their websites and publicly available data. Thus, some information might be missing from the categorisation.

⁵ A human actor gives feedback.

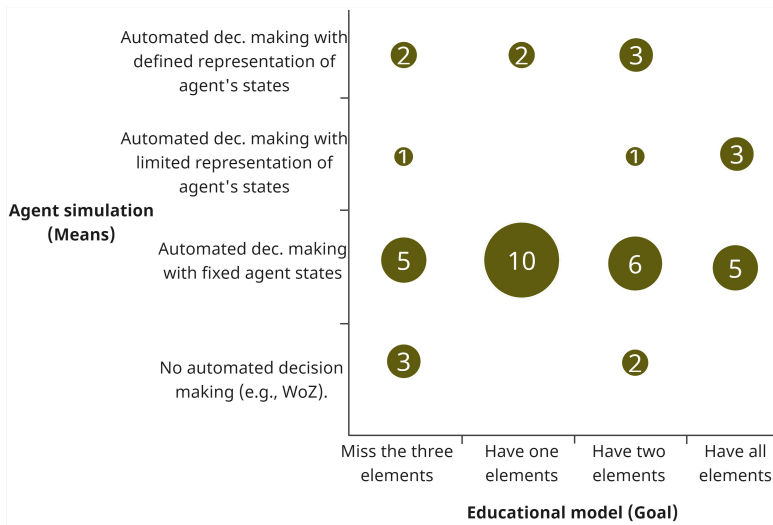


Figure 2.6: Categorisation of the 43 training systems based on criteria related to the agent simulation model and the educational model, as detailed in Tables 2.5 and 2.6. The numbers inside the circles represent the count of systems that correspond to each category.

The table also shows that all five components are present in the examined training systems, suggesting their relevance in such contexts. Each concept discussed in the categorisation table (Table 2.5) also appears in at least two of the examined systems, further validating the relevance of the ideas. Figure 2.6 shows the extent to which the 43 systems cover the agent simulation model and the educational model. We assessed the agent simulation based on the complexity of decision-making and the agent model, and we assessed the educational model based on the presence of three elements: feedback, multiple instructional principles, and the use of a pedagogical agent. None of the systems achieved the highest scores in both dimensions, while some scored highly in mainly a single dimension, suggesting a focus on either one over the other⁶. Furthermore, the training systems analysed span various topics, from job interviews to suicide prevention and negotiations. This broad representation emphasises the general applicability of the five components across various training types, thus supporting ARTES's generalisability.

2.4.3 TOOLS SUPPORT FOR ARTES COMPONENTS

Our previous discussion examined the link between ARTES components and other architectures conceptually. Now, we turn our attention to the implementation support for ARTES components, i.e., do existing tools support the implementation of ARTES components? To answer this question, we linked ARTES to three development platforms for creating virtual humans: GRETA [307], the Virtual Human Toolkit (VHToolkit) [155, 156], and Agents United [36]. While these platforms are mainly designed to develop virtual humans, they

⁶Similar to the mapping of architectures, systems are typically designed to suit their specific contexts. Thus, our analysis does not assess their overall quality; instead, it aims to show how these systems incorporated ARTES's components, highlighting its relevance and modularity of the architecture.

are also suitable for agent-based social skills training systems. Table 2.8 depicts which tools can be utilized to build ARTES agent simulation components in these three platforms. The linking shows that the tools can support the implementation of ARTES simulation components. For a comprehensive overview of virtual human tools and their applications, we refer to the chapter by Hartholt and Mozgai [154].

Table 2.8: Mapping virtual human platforms and their tools to ARTES’s agent simulation components.

ARTES Components	GRETA	VHToolkit	Agents United
Simulation Environment	FAP-BAP Player Ogre3D Unity Robotics	Unity	Unity Robotics
Speech Recognition	PureData	MS SAPI Google Voice PocketSphinx	N/A
Multimodal Perception and Understanding	SSI OpenFace OpenSmile	MultiSense	Holistic Behaviour Analysis Framework AWARE universAAL
NLU	Emotional Mind- Dialog Manager	NPCEditor	WOOL
Agent Model & Action Decision Making	Intent Planner TRINDIKIT MIDAS	NPCEditor Scripting	Topic Selection Engine Dialogue Game Execution Flipper WOOL
NLG	Emotional Mind- Dialog Manager	NPCEditor	WOOL
Speech Generation	MaryTTS CereVoice	Festival MS SAPI Rhetorical CereVoice	MaryTTS MS SAPI
Nonverbal Behaviour Generation	Behavior Planner	NVBG	Flipper
Nonverbal Behaviour Realizer	Behavior Realizer	SmartBody Unity BML Realizer	GRETA Behavior Realizer ASAP Realizer
Animation System	OpenGL Library Unity Mecanim	Unity Mecanim	Unity Mecanim

Regarding ARTES’s educational model, we can map ITS-based authoring tools directly to it, as ARTES’s educational model also relies on ITS main components (domain knowledge, learner knowledge, and tutor). ITS-centric examples include GIFT [350] and AutoTutor [140]. Integration of authoring tools with virtual human platforms would align with the main two parts of ARTES: the virtual agent simulation, and the ITS-based educational model. Such integration is possible, as shown by merging the VHToolkit with AutoTutor [358], and the VHToolkit with GIFT [57]. In the future, ARTES could be evaluated through an example implementation using such tools or by implementing a prototype from scratch, allowing us to evaluate other quality attributes such as performance and usability [1].

2.5 FURTHER RESEARCH DIRECTIONS

What research directions still warrant further exploration? We identified nine of them, which we will discuss here one by one.

Knowledge Transfer. Knowledge transfer in education relates to learners' ability to generalise the gained knowledge to other contexts [29]. In training systems, this translates to transferring knowledge gained from agent simulations to real-world situations. Various concerns arise regarding knowledge transfer in training systems, notably: (a) the inaccuracy of reflecting complex, real-world interactions; (b) the simulated situation being context-dependent; and (c) the lack of motivation to learn and transfer skills to other contexts. Despite the transfer's importance, most examined systems overlooked the assessment of real-life knowledge. In their review, Bosman et al. [48] reported one system out of twelve that tested for knowledge transfer, attributed to the cost and expertise needed to conduct it, especially as computer scientists in academia mainly created these systems. Naturally, testing for such tangible outcomes should accompany adherence to defined design guidelines and recommendations for fostering knowledge transfer (e.g., Craig and Schroeder [92]). Regarding the provided feedback, away from the agent simulation, it is worth noting the critical role of instructions specificity level for knowledge transfer. Eiriksdottir and Catrambone [110], for instance, examined educational instructions and found that participants receiving detailed, step-by-step instructions performed well initially but received worse learning and transfer results later. On the other hand, they found that general instructions aided people in establishing a more adaptable reference framework for new tasks. This finding supports the need for knowledge transfer testing in agent-based training systems, as good initial performance results may not necessarily translate into successful knowledge transfer.

Balance between realism and educational gains. High-fidelity simulation does not necessarily correspond to better learning outcomes [70]. For training systems, the realism of a simulation could cover elements such as the rendered agent and environment, the agent's behaviour, dialogues, and sound. Gallagher et al. [134] argued that the essential question is whether the simulation trains the appropriate skills, regardless of the technological fidelity. At the same time, realism in simulation can impact attributes such as the agent's likability [32] and the satisfaction of the agent [308], which could, in turn, influence educational gains. Thus, the challenge lies in identifying essential simulated elements that can fulfil learning goals. More specifically, how to find cues or behaviours that are sufficiently realistic to train a specific skill and responses and meet the training objectives. For instance, adjusting eye gaze intensity can help manage social phobia in virtual reality exposure therapy [59], or altering depth perception can address fear of heights [230]. As for establishing a balance between training objectives and simulation fidelity, we took steps towards it by linking learning theories to ARTES components, which would contribute to determining the level of development required for each component, guided by the targeted skills and relevant theories.

Explainability of training systems decisions. The nature of training systems requires two types of explanations for learners: one addressing the virtual agent's behaviour, and the other offering insights into the learner's choices as part of the educational process. There is a need to integrate these two aspects within the context of training systems. Doing that would help generate and present reliable and valid feedback that explains the

agent's actions and the learner's decisions. When it comes to the explainability of the agent, we need to explain the agent's actions, their consequences, and why the scenario unfolded in the way it did. This is often referred to as explainable agency, and education is one of its drives for demanding explainability [17]. This explainability typically unfolds in three phases: generating the explanation regarding the action, communicating the reason to the learner, and assessing the explanation reception [281]- the latter equivalent to experiments in the case of training systems. The other type of explanation is regarding the learner's performance in training. One such work in this domain is where Khosravi et al. [199] defined a framework for Explainable AI in education, which are six dimensions that education systems developers and researchers should consider.

Ethical implications. Using interactive agents in training systems could raise stereotypes and biases. Rivera-Gutierrez et al. [314] experimented with six virtual patients varying in gender and skin colour, revealing a disparity in correct diagnoses among the different characters. The authors speculated this could be due to the transfer of prejudices from real life to the simulation. Conversely, this transfer could potentially develop biases based on the learner experience within the training system. For instance, if a character with certain traits behaves in a particular way, it could inadvertently reinforce discriminatory attitudes towards such traits or populations. Moreover, modelling "appropriate" behaviour in agent-based training systems raises two main concerns: regarding (1) inaccurate design and (2) potential malicious control of this behaviour. If an agent's behaviour is poorly designed, simulations might not accurately represent human interactions. This could lead to ineffective training, giving learners a false sense of competence. As for who is controlling appropriate behaviour, training systems could be used to shape and moralise people on specific behaviours. Similarly, ill-intentioned learner could misuse the skills learned from training systems to cause harm, such as exacerbating bullying. Regarding affective privacy in virtual agents, Hudlicka [174] highlighted three ethical concerns, which are apparent in a training system that tries to change the user's emotions, e.g., social anxiety training. The concerns are affective privacy, or the right to keep one's emotions private; emotion induction, the act of evoking and changing specific emotions in a person; and the formation of virtual relationships through (over)trust and attachment. If these systems are inadequately designed, for instance, by evoking inappropriate emotions in the learners, they could have negative consequences that exacerbate the issues. Other ethical questions for training systems could arise while analysing and storing personal data and regarding consent and autonomy for such simulation training. A solution that could address this is to develop guidelines and checklists as a community, which can provide a structure for design validation. Furthermore, we should involve multiple stakeholders in the evaluation of training systems to get a variety of perspectives on system validity and preferences (e.g., Botezatu et al. [52]), and by letting them participate in the design process (e.g., Kleinsmith et al. [206], Kron et al. [213]).

Large Language Models (LLMs). These models, such as GPT [62], can be considered as another option for simulating a person in a training scenario. Figure 2.7 shows an interaction with ChatGPT acting as a child contacting a children's helpline. In this case, the conversations appear convincing and reasonably coherent. However, there are currently several challenges to the direct application of LLM in social skills training systems. For example, it is less clear how to explain agents' cognitions in the absence of a human-

ChatGPT:

Sarah: I guess I just needed someone to talk to. I haven't really told anyone about this, not even my parents. I just feel so alone. Maybe you can give me some advice or tips on how to deal with them or make it stop? Or just help me understand why they're doing this?

Human:

Unfortunately, I can't do that directly. But we can look for a different solution together. What do you think of that?

ChatGPT:

Sarah: Okay, I'd like that. I just want things to get better, you know? If we could find a way for me to cope or maybe even get them to stop, I'd be so grateful. Anything to make the days easier.

Human:

Do you know someone who would listen to you about this problem? Maybe a teacher or your parents.

ChatGPT:

Sarah: I haven't told my parents because I don't want them to worry. But there's this one teacher, Mrs. Roberts. She's always been nice to me. Maybe I can talk to her about it? I'm just afraid that if I tell a teacher, the bullies will find out and make things even worse for me.

Human:

Yes I think it will be a good idea to talk to Mrs. Roberts. when can you talk to her?

Figure 2.7: A chat from ChatGPT (GPT-4). We prompt the tool to act as a bullied child contacting a children's helpline. The prompt was "*Engage in roleplay. You are a child (Sarah) who was bullied at school and chatting with a child helpline. I will play the volunteer at the helpline. Let's start: "I am John, I am here to listen to you. What do you like to talk about?"*"

understandable agent model. Additionally, LLMs are reported to suffer from bias issues [277] and carry a risk of misinformation in the agent reply [392], which could deviate the training from its objectives. One possible approach to utilizing LLMs is to simulate some parts of the agent's thinking. These models generally generate fitting responses, but may occasionally deliver incorrect information [194]. Therefore, exploring hybrid solutions could be beneficial, blending rule-based agent models and LLMs to generate appropriate agent responses. This could involve generating parts of the agent's thinking, like their beliefs, intentions, and actions, to be fed into a framework [20]. Another outlook involves incorporating LLMs into the tutor model to supplement or potentially replace certain components. Despite their limitations, LLMs could automate sections of the educational process and reply to unexpected learner inputs. For Lila's case, three examples can illustrate the potential applications of LLM-based tutor subcomponents. First, a pedagogical agent could serve as a social actor, addressing Lila's questions about communication protocols and counselling in general. Second, the responses of the virtual child character could be personalised based on Lila's demonstrated proficiency, thus reflecting instructional principles. Lastly, feedback for Lila could be generated by comparing her conversation logs with the virtual child to an ideal conversation log.

Crowdsourcing solutions. As previously mentioned, data required to create a virtual human or a tutoring model is not always accessible or adequately expansive for every training system domain. One solution to such a challenge is crowdsourcing. In theory,

crowdsourcing could aid in collecting data required for a data-driven virtual agent. However, in most training systems, a domain expert is needed to feed and evaluate such data, thereby limiting the pool of workers. Nevertheless, crowdsourcing has proven helpful for training systems; for example, Rossen and Lok [321] proposed an approach to develop a conversation with a virtual agent rapidly. This approach allows a domain expert to start an agent with an initial set of questions and responses, which will then interact with novices recruited through crowdsourcing. The domain expert then refines the agent with information from novice interactions. This iterative procedure continues until the agent's performance reaches a satisfactory level. Such an approach could decrease the cognitive load and workload on the domain expert and yield a broader range of responses. Another approach that focuses on simulating different actions in a situation was proposed by Feng et al. [118]. In that study, the workers were asked to complete a situation story and how it could unfold. These stories were fed to data-driven models, and the workers evaluated the outputs iteratively. Although this approach reveals actions and consequences within a situation, it falls short in modelling conversations—meaning the agents' dialogues are not directly trained. Still, it can be an initial step to creating a good story for future dialogues. On the agent model level, crowd workers could contribute to defining the agent's states, based on the current state of the conversation.

Multi-agent scenarios. We, with ARTES, only focused on social scenarios involving a single virtual agent. However, training social skills might require the presence of multiple actors, and thus, there is a need to model more than one agent to simulate a situation. Depending on a scenario's context and learning objectives, agents could have individual cognition or share a collective cognitive model. For example, a virtual nurse and a virtual patient could be simulated to train a student doctor [213]. Furthermore, multiple agents can be used with social cognitive theory to model agents with differing perspectives on a topic that a learner could learn from. For example, learners can explore the ethical implications of implementing a new two-factor authentication policy by engaging with agents that prioritize different ethical principles, such as non-maleficence (not harming people) or autonomy (giving people choice) [12]. It would be interesting to explore the dynamics of multiple agents and how their shared cognition could enhance the learner's training. Taking this concept further, the simulation system could train more than one learner in a collaborative learning environment. An example of this approach is utilising a virtual agent moderating conflict resolution between two learners [111].

Research methods. Prior research has discussed the lack of standardised terminology and standardised evaluation of learning outcomes for virtual patients [30] and training systems in general [48]. As a next step, the community could establish strategies to enhance the reproducibility of results. There are two strategies to foster the future reproducibility of the system's positive learning effects. The first one focuses on specifying the material used, for example, a model to illustrate the internal functioning of the agent. However, crafting an accurate model presents its challenges, as these models represent an abstraction of reality. Thus, an alternative strategy is to record people's impressions and test whether these can be replicated as a blackbox approach. One of the initiatives for this strategy is the Artificial Social Agents (ASA) questionnaire [124], which allows researchers to measure a broad number of interaction impressions, such as the believability, enjoyability, and social presence of an agent [122].

Procedurally generated scenarios. Regarding social interaction training systems, there are many intersections with serious games – games designed for objectives beyond entertainment. They both have goals, objectives, and, usually, virtual characters. Incorporating gamification elements can increase engagement and overall learning outcomes [173, 325]. This can be achieved by using scores as rewards [135, 184], or using badges [410]. One direction forward for training systems is incorporating procedurally generated dialogue, where the storyline, characters and their proprieties (e.g., traits and goals) dynamically evolve as the scenario progresses. For example, the game “Tech Support: Error Unknown” [348], which, despite not being the main focus, does exhibit some training elements. Here, the player acts as a technical support representative who people contact for support, where the customer’s support requests are procedurally generated. Having a similar procedural narrative could introduce an element of randomness to scenarios in training systems, thus mitigating repetitiveness and predictability.

2.6 FINAL REMARKS

We believe ARTES is a valuable foundation for researchers and developers in agent-based social skills training systems. This paper’s contribution lies in the proposal of this versatile architecture and the examination of its key components, comparing them to similar domains and linking them to learning theories. Understanding the complexities and roles of these vital components, we hope, could enhance the efficiency, cost-effectiveness, adaptability, and impact of training systems. Therefore, our work serves as a stepping-stone to bridging the gap between virtual human simulations and effective educational learning.

3

3

SIMULATION-BASED TRAINING FOR CHILD HELPLINE COUNSELLORS: IMPACT OF SIMULATION, FEEDBACK, AND REFLECTION IN A RANDOMISED CONTROLLED TRIAL

Child helplines require well-trained counsellors to support children in need. Such training typically involves role-playing, which is effective but costly and difficult to organise at scale. A promising addition, therefore, is to offer a simulation-based training, where trainees interact with a conversational agent that mimics a child contacting the helpline. However, interacting with a simulation alone may not be sufficient, as augmenting it with feedback and reflection can provide more guided learning. The learning effects of sequentially adding these elements—simulation, feedback, and reflection—to training systems are not yet fully understood. In this paper, we extend Lilobot, a conversational BDI-based virtual child, to investigate the effects of these elements on learning outcomes. In a randomised controlled online trial ($N = 346$), participants were randomly assigned to one of four conditions: no intervention, simulation only, simulation with feedback, or simulation with feedback and reflection. Participants interacted with the interventions in five sessions over fourteen days. Compared with the no-intervention condition, simulation training improved participants' task knowledge and reflective-writing capability. Adding feedback led to further gains in knowledge and conversational outcomes (performance). Also, adding reflection on top of this

📄 **Al Owayyed, M.**, Tielman, M. L., and Brinkman, W.P.. Simulation-Based Training for Child Helpline Counsellors: Impact of Simulation, Feedback, and Reflection in a Randomised Controlled Trial. *Manuscript is under review.*

further increased reflective-writing capability, though it dampened conversational outcomes improvements. These findings clarify how simulation, feedback, and reflection each contribute to learning in early-stage counsellor training.

3.1 INTRODUCTION

Child helplines provide support and advice to adolescents and children on various issues, including mental health, sexuality, bullying, and family problems. Typically, children can access these services by interacting with a counsellor via phone or chat. The international child helpline network reported that its more than 150 member helplines around the world received over 6 million contacts in 2023, more than 2 million of which were counselling contacts [176]. The British Childline receives a contact through its various channels every 45 seconds on average [284]. Similarly, the Dutch child helpline, De Kindertelefoon, received an average of 1000 contacts per day in 2023 [355] and 900 per day in 2024 [355]. To cope with this demand, De Kindertelefoon trained more than 640 new volunteers over these two years.

For such interactions to be effective, counsellors are trained to apply counselling protocols and theories. However, training counsellors typically requires multiple sessions rather than a single interaction. For example, a one-day training workshop on Motivational Interviewing is not sufficient for developing sustained skills or improving client outcomes [260, 262]. Training counsellors' skills often includes a role-playing setting, where a counsellor takes on the role of the person contacting the helpline [130, 220]. However, reinforcing these skills through role-playing requires repeated practice [240], which is costly, difficult to coordinate, and time-intensive to scale. For instance, both trainers and trainees must be present simultaneously to conduct role-play sessions. An online simulation training tool has the potential to overcome these limitations when incorporated with traditional training. These tools could provide a controlled, readily available, and affordable training environment [143, 247, 313].

Several conversational agent (chatbot) simulations have been developed with similar aims, such as training on interrogation techniques [64], diagnostic skills [166], or sales-talk skills [272]. Nevertheless, limited research has focused on training helpline counsellors using conversational agents, especially for training counsellors at children's helplines. Demasi et al. [99] proposed a chatbot to train suicide prevention hotline counsellors, evaluating how participants perceive the chatbot across several dimensions (e.g., coherence and usefulness). Another example is ClientBot, which is designed to train basic counselling skills, including open questions and reflective statements [364]. In the context of children's helplines, Veldhuizen-Ochodničanová et al. [382] presented a serious game to train counsellors in skills such as empathy and communication. This system follows a choice-based script in which trainees select from predefined replies, rather than generating their own. As a result, it provides less practice in formulating responses from scratch compared to open-ended interaction.

Before helplines consider integrating such a simulation tool into their curriculum, they need to understand what it should include and how effective it will be. Specifically, this involves identifying which components make a simulation training system effective and how different elements can enhance trainees' skills and performance. The ARTES architecture [8] identifies agent simulation and educational components as key components

of social skills training systems. Using these two key components, this paper reports an experiment examining their incremental addition to a training tool and the resulting effects on the learning process and outcomes. The elements include the simulation alone, the simulation with feedback, and the simulation with feedback and engagement through reflections. We conducted a randomised controlled online trial ($N = 346$), in which potential new trainees were assigned to a control condition or one of three versions of the tool across five sessions over fourteen days. We assessed their learning outcomes, performance, and perceptions, as well as behavioural patterns related to different feedback levels and how these may influence learning.

3.2 HYPOTHESES AND RESEARCH QUESTIONS

3.2.1 SIMULATING HUMANS FOR SOCIAL SKILLS TRAINING

Training with simulations can help people develop important skills. Social skills training systems have been applied in various domains, including aggression de-escalation [49], medical education [388], and job interviews [15]. A recent literature review [103] found that most of these systems report positive outcomes on measures such as improved conversational skills. However, these systems are often evaluated as complete packages that combine simulation with additional elements such as feedback or guidance, rather than isolating the role of the simulation itself. This distinction is important because it clarifies whether improvements stem from the simulation itself or from elements that enrich and build upon it. Still, studies on simulation-only training systems found positive outcomes, particularly in how participants perceive and engage with the simulation [108, 239].

Interacting with a simulation can align with the behaviourist perspective [389], which focuses on reinforcing desired learner behaviours through repetition and feedback. At the same time, simulations engage trainees at higher levels of Bloom's taxonomy [43] (applying and analysing) because they require active decision-making, testing of different possibilities, and adaptation of actions based on simulated outcomes rather than simple recall of information. This active engagement encourages trainees to think critically about their actions, evaluate strategies, and reflect on their performance. Such deeper application can foster metacognitive abilities [3, 297], including understanding and reflection [205, 207]. Therefore, our first hypothesis is as follows.

Hypothesis 1 (H_1). *Compared to those who only received the theoretical explanation, people who interact with a simulation demonstrate (1-a) a higher increase in task knowledge and (1-b) a higher reflection capability.*

3.2.2 FEEDBACK IN SIMULATION-BASED TRAINING SYSTEMS

Learning solely through simulation aligns with a behaviourist perspective, which focuses on enforcing desired learner behaviour (e.g., through repetition) [389]. Adding guidance to the simulation introduces a cognitivist perspective, emphasising the development of knowledge through mental processes [113]. Blending these perspectives can enhance the learning process [113]. Also, incorporating tutoring techniques into the simulation could lead to a more guided learning experience [204].

Delivering feedback solely through the simulation may not be sufficient. For example, feedback embedded in the agent's behaviour (e.g., leaving the conversation) helped par-

Participants recognise what had happened and whether mistakes were made, but they still reported a need for more clarification about what went wrong [283]. Several simulation-based training systems have incorporated different forms of feedback that have been shown to enhance trainees' knowledge and performance. This includes delayed feedback provided after the interaction (e.g., [15]) and immediate feedback delivered during the interaction (e.g., [273]).

Feedback has been shown to improve both cognitive and motivational outcomes [396]. One study that examined this addition is the MACH system, a job interview training simulation [171]. Participants were assigned to one of three conditions: watching an educational video (control), using MACH without feedback, or using MACH with feedback. Human experts rated participants' performance, and results showed higher improvement in the feedback condition compared to both the simulation-only and the control conditions, while no difference was reported between the simulation-only and control conditions. Similar results were found with ClientBot [364], where adding real-time feedback to a simulated patient improved participants' performance compared to using the simulation alone. However, the feedback focused on the trainee's shown skills and on suggesting the next skill to use, rather than on the overall counselling process or session structure (see the task-level and process-level discussion in Section 3.2.4). Although feedback is widely used in agent-based social skills training systems, its design and impact on learning in helpline training simulations are not yet well understood. This is partly because helpline training requires simulating a conversational agent using text only. Altogether, we propose the following second hypothesis:

Hypothesis 2 (H_2). *Compared to a simulation only, people who interact with a feedback-integrated simulation demonstrate (2-a) a higher increase in task knowledge, (2-b) a higher reflection capability and (2-c) higher conversational outcomes (performance) over time.*

3.2.3 REFLECTION MOMENTS IN SIMULATION-BASED TRAINING SYSTEMS

Adding reflection moments to training, in the form of self-reflection, could enhance the learning process. This is supported by Kolb's Learning Cycle, which shows the importance of reflection in experiential learning [210]. Several researchers have established links between effective learning and reflection (e.g., [227, 390]). Incorporating self-reflection moments into a training system adds another learning perspective, constructivism, where learners construct their own knowledge [113].

Hoang et al. [168] showed that incorporating self-reflection prompts into an online clinical skills training improved learners' knowledge acquisition, interpersonal skills, and performance compared to training without self-reflection prompts. Additionally, in a medical education context, Maqsood et al. [244] found that integrating self-reflection with feedback led to higher learning outcomes than reflection alone, highlighting the importance of combining both. Despite this, self-reflection is not commonly integrated into training simulations.

As mentioned earlier, training systems can improve reflection. Reflection can be further enhanced through repeated opportunities, helping learners develop stronger reflective abilities over time. Therefore, we propose the third hypothesis:

Hypothesis 3 (H_3). *Compared to a feedback-integrated simulation, people who interact with a feedback-simulation extended with reflection demonstrate (3-a) a higher increase in task knowledge, (3-b) a higher reflection capability and (3-c) higher conversational outcomes (performance) over time.*

Building on H_1 – H_3 , we expect a cumulative effect, such that the sequential incorporation of simulation, feedback, and reflection leads to progressively higher learning outcomes. Therefore, we propose the following overall hypothesis:

Hypothesis 4 (H_4). *The subsequent incorporation of simulation, feedback, and reflection shows an overall (4-a) higher increase in task knowledge, (4-b) higher reflection capability, and (4-c) higher conversational outcomes (performance) over time.*

3

3.2.4 HOW TRAINEES USE TASK- AND PROCESS-LEVEL FEEDBACK

Although feedback and reflection are useful for training, they can also overwhelm trainees if not carefully designed. According to Hattie and Timperley [158], feedback can occur at four levels: *task* (correctness of work), *process* (strategies and methods), *self-regulation* (self-monitoring and control), and *self-level* (personal evaluation). In training contexts, *task-level* feedback supports improvement of immediate actions, whereas *process-level* feedback supports understanding of the underlying strategy and structure. These two levels are particularly important because they help students refine tasks and better understand the processes involved. However, cognitive load theory suggests that too much information can overwhelm cognitive resources [360], which could happen with feedback [158]. Despite this, little is known about which types of feedback learners prefer or which can support learning in training simulations.

Multiple mechanisms influence the levels of feedback learners require during training. Gagné's Hierarchy of Learning presents a bottom-up progression in which learners first acquire prerequisite capabilities, such as stimulus-response, before building on more complex skills, such as problem solving [133]. From this framework, one might infer that feedback should initially focus on lower-level task performance and then shift toward process-level feedback as learners advance. At the same time, some instructional frameworks, such as advanced organisers and meaningful learning theory [24], suggest beginning with higher-level understanding (similar to process-level feedback) before focusing on more detailed information (task-level). Existing recommendations and studies on effective feedback (e.g., [158, 341]) often recommend combining the different levels of feedback.

Given these perspectives, people may engage with different levels of feedback in different ways, and specific usage patterns may be associated with better performance. For example, if trainees prefer a certain type of feedback, introducing that type early in the training may be more effective. Previous studies in social skills training have examined the effects of feedback timing (immediate vs. delayed) [84], but not its levels. Here, we therefore propose the following explorative research questions:

RQ1: How do people use different levels of feedback over time?

RQ2: How does it influence their task knowledge, reflection capability, and conversational outcomes?

3.3 METHOD

We conducted a mixed-method experiment with 346 participants: a between-subjects design to test the hypotheses and a within-subjects design to analyse the research questions. All participants watched a video explaining the communication protocol. They were then randomly assigned to one of four conditions: no additional intervention (*C*, i.e., Control); interaction with the simulation only (*S*); simulation with feedback (*SF*); or simulation with feedback and reflection (*SFR*). Participants in conditions *S*, *SF*, and *SFR* interacted with the training tool in five sessions, each conducted every third day (i.e., with two days between sessions), to assess the effects of interaction over time. Before the experiment, we conducted a pilot study with 16 participants to refine the tool's instructions and reflection questions, as well as to improve how the system recognised participants' inputs. These participants completed all five sessions, and their data were not included in the final analysis. The study was approved by the university's Human Research Ethics Committee (ID: 3946) and was pre-registered prior to data collection ¹.

3

3.3.1 PARTICIPANTS

In total, we recruited 738 participants through the online platform Prolific, of whom 346 completed the experiment and comprised the final sample. The inclusion criteria were fluency in English, a Prolific approval rate of at least 95% with five or more completed studies, and holding at least a Bachelor's degree—a prerequisite for counsellors starting training at the Dutch children's helpline. A summary of the demographics per condition is shown in Table 3.1. Figure 3.1 illustrates the participant flow and exclusion reasons per condition. Demographics of excluded participants are reported in Appendix Table C.2. Overall, there were no clear differences in demographic characteristics between excluded participants and the final sample, except that excluded participants were slightly more likely to have children; however, this difference had a small effect size (see the Appendix for details). All participants were compensated at Prolific's minimum payment guideline (£6 per hour) for the first four sessions, and £10 per hour for the final session.

To estimate the required sample size, we conducted a priori power analysis (using Frequentist methods). This resulted in at least 64 participants per condition to achieve statistical significance at the 95% confidence level with 0.8 power, assuming a medium effect size for an independent *t*-test [88]. For the exploratory research questions (i.e., those concerning condition *SFR*), we conducted 1,000 simulations of the model using synthetic data to estimate the necessary sample size. These simulations, based on the models we planned to test, suggested that 150 participants would be sufficient to achieve a desired power of 0.8 with 95% confidence. Therefore, we assigned participants using a weighted randomisation (43% *SFR*; 19% each for the other conditions) to ensure target sample sizes of at least 150 in *SFR* and 64 in each of the remaining conditions.

¹<https://osf.io/j9k4u>

Table 3.1: Summary of participant demographics by condition (control, simulation only, simulation and feedback, simulation, feedback and reflection). Full details are in Table C.1 in the appendices.

Demo-graphic	Control (n=66)	Simulation (n=66)	Simulation and feedback (n=64)	Simulation, feedback, and reflection (n=150)
Age Average	$M = 31.41, SD = 9.58$	$M = 31.50, SD = 9.13$	$M = 31.75, SD = 8.59$	$M = 31.90, SD = 10.24$
Gender	Men: 30 (45.5%) Women: 36 (54.5%)	Men: 28 (42.4%) Women: 38 (57.6%)	Men: 32 (50%) Women: 32 (50%)	Men: 62 (41.3%) Women: 81 (54%) Non-binary: 4 (2.7%) Prefer not to say: 3 (2%)
Educa-tion	Community/Technical: 3 (4.5%) Bachelor's: 44 (66.7%) Master's: 15 (22.7%) PhD: 4 (6.1%)	Community/Technical: 3 (4.5%) Bachelor's: 49 (74.2%) Master's: 12 (18.2%) PhD: 2 (3%)	Community/Technical: 6 (9.4%) Bachelor's: 41 (64.1%) Master's: 16 (25%) PhD: 1 (1.6%)	Community/Technical: 13 (8.7%) Bachelor's: 99 (66%) Master's: 31 (20.7%) PhD: 7 (4.7%)
Has Children	Yes: 30 (45.5%) No: 34 (51.5%) Prefer not to say: 2 (3%)	Yes: 32 (48.5%) No: 30 (45.5%) Prefer not to say: 4 (6.1%)	Yes: 36 (56.2%) No: 25 (39.1%) Prefer not to say: 3 (4.7%)	Yes: 82 (54.7%) No: 62 (41.3%) Prefer not to say: 6 (4%)
Experi-enced Bullying	Yes: 60 (90.9%) No: 4 (6.1%) Prefer not to say: 2 (3%)	Yes: 60 (90.9%) No: 6 (9.1%)	Yes: 57 (89.1%) No: 6 (9.4%) Prefer not to say: 1 (1.6%)	Yes: 132 (88%) No: 15 (10%) Prefer not to say: 3 (2%)
Chatbot Experi-ence	NA	Once or twice: 4 (6.1%) Occasionally: 25 (37.9%) Frequently: 37 (56.1%)	Once or twice: 12 (18.8%) Occasionally: 18 (28.1%) Frequently: 34 (53.1%)	Never: 1 (0.7%) Once or twice: 13 (8.7%) Occasionally: 54 (36%) Frequently: 82 (54.7%)

3.3.2 TRAINING SYSTEM DESIGN

To examine the impact of subsequently adding simulation, feedback, and reflection on learning outcomes, we updated Lilobot's simulation to include these components. Therefore, we ended up with three versions of the system: Lilobot alone (condition *S*), Lilobot with feedback (*SF*), and Lilobot with feedback and reflection (*SFR*). In this section, we address the following questions: How does each component work? And what does the overall structure of these components within a learning session look like?

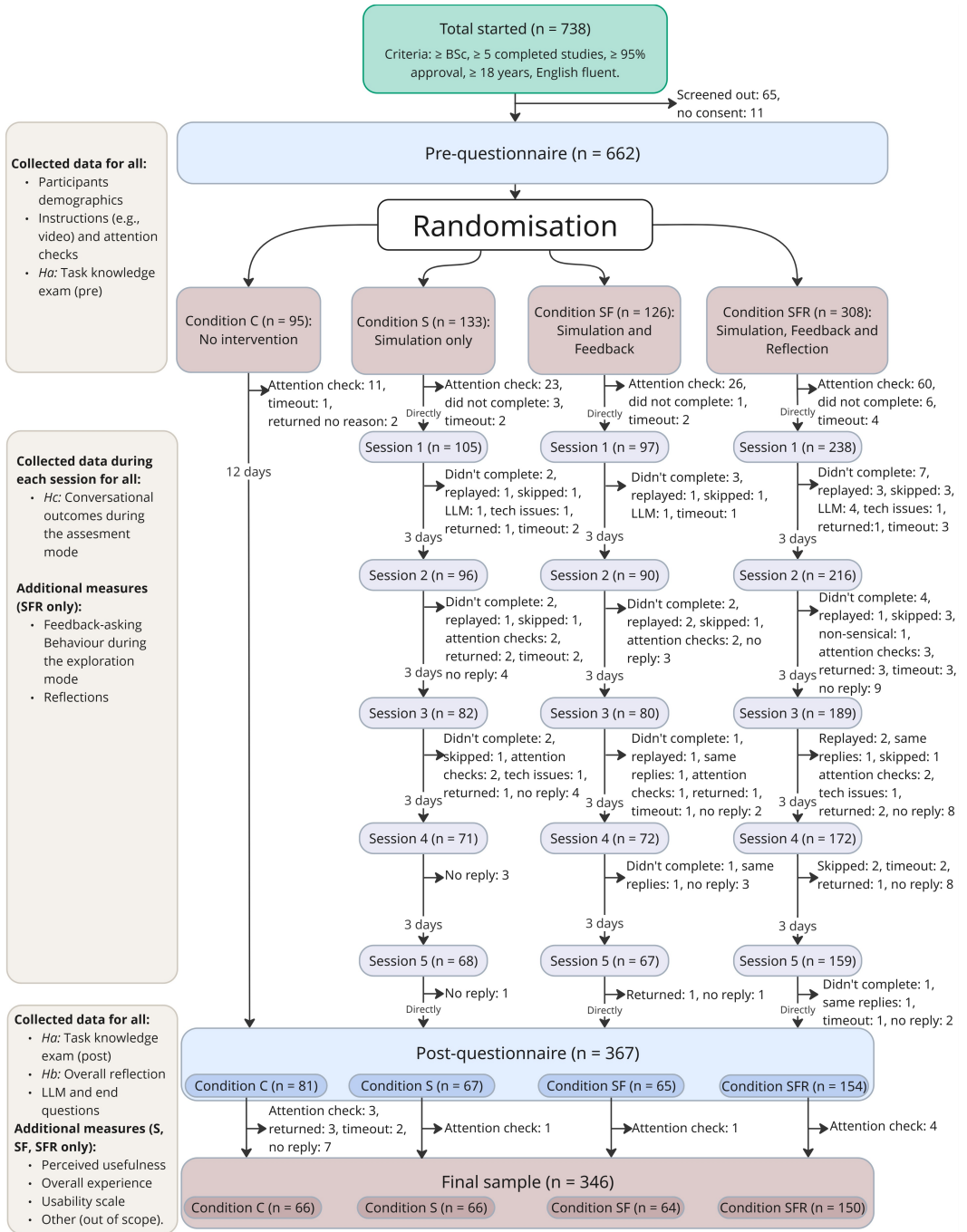


Figure 3.1: A Flowchart of participants' progression through the experiment, including recruitment, condition assignment, exclusions, session activities, and measures. Each session followed the structure in Figure 3.3. Invitations for the next session were sent three days after each session.

SIMULATION THROUGH LILOBOT

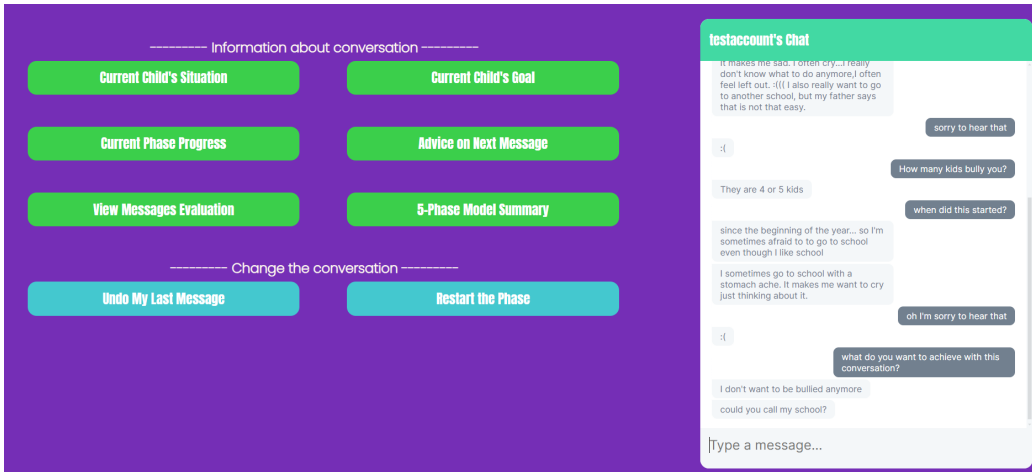


Figure 3.2: Screenshot of the training system. The right panel displays a chat simulation with Lilobot during the transition from phase 2 to phase 3 of the five-phase model, while the left panel provides immediate feedback options for the participant to use.

Lilobot is a virtual child for training counsellors-in-training at helplines [145]. With this tool, a trainee interacts with a conversational agent mimicking a virtual child, whom we call Lilobot. Lilobot simulates children reaching out to a helpline to discuss issues, such as being bullied at school, through an open-ended chat-based channel. The trainee aims to counsel Lilobot by applying communication protocols, such as the five-phase model [344], which was developed for children’s helplines. The phases of this model are building rapport, clarifying the child’s story, defining the conversation goal, working toward the goal, and wrapping up the conversation [344]. This protocol guides the conversation while ensuring that it remains child-focused.

Lilobot’s decision-making process is based on a rule-based beliefs-desires-intentions (BDI) model, which has been used to simulate various interactive agents for training purposes (e.g., [15, 64]). In brief, the agent holds several beliefs about itself and the world (e.g., “Lilobot thinks the trainee can be trusted”). These beliefs are influenced by a predefined mapping of the participant’s inputs. For example, if the participant tells Lilobot, “If I understand you correctly, you are being bullied, and you want it to stop, is that correct?”, Lilobot’s trust belief in the trainee increases. Lilobot’s desires represent the child’s goals during the conversation (e.g., “Lilobot wants to talk about their problem”), while intentions reflect the currently active desire. Lilobot’s current beliefs trigger these active desires. The agent then selects an appropriate response based on the participant’s most recent input and the active desire from the knowledge base. The conversation can be concluded in two ways: either when the participant successfully applies the five-phase model, or Lilobot perceives the interaction as unhelpful or intimidating. Figure 3.2 illustrates an example interaction in which a participant moves from the end of phase two (clarifying the child’s story) to

phase three (setting up the conversation goal). A detailed description of the simulation is provided in [145].

Table 3.2: Scenario settings of twelve virtual children, including the bullying setting, the wrong goal attributed to the child, and the corresponding child names.

Setting	Challenging Goal			No Goal	Non-challenging Goal
	Call School/ Parents	Stop Going/ Playing	Hit Back/ Revenge		
School break	Lilo				Miles
School music					Maya
School class		Daisy			
Sports club			Tim	Jim	
At home				Lyla	
Neighborhood	Axel				Ori
Online gaming		Kai			
Social media			Ava		
Playdate				Dania	

Grundmann et al. [145] introduced only a single scenario in Lilobot: Lilo being bullied at school and asking the trainee (i.e., participant) to call their school to stop the bullying. In the current study, we extended the training to twelve different scenarios, all centred on bullying but varying along two dimensions: the setting (e.g., school, home, neighbourhood, online) and the child’s initial goal in the conversation (e.g., asking the participant to call their school, not knowing what to do, or seeking revenge). By introducing multiple scenarios, we hoped to improve the generalisability of both the experimental design and the training outcomes. Relying on a single scenario risks overfitting the design and participants’ learning to a specific situation. Moreover, exposure to varied scenarios limits the potential learning effects of repetition, while diversity enhances training quality. We also introduced scenarios with challenging goals (e.g., wanting to hit back), non-challenging goals (e.g., the child is ready to talk), and no goal (e.g., not knowing what to do), so that participants would not develop biases towards a particular scenario during training. The scenarios focused primarily on bullying because it is multifaceted [195], can take many forms [68], and is a frequent issue in helpline contexts [249], while still allowing for grounded and comparable solutions across cases. Table 3.2 presents the twelve virtual child scenario settings and their names.

To generate the virtual children’s responses, we first provided GPT-4o with utterances from the original scenario in [145], along with a new scenario setup and child persona, and asked it to produce corresponding utterances for the new scenario. These dialogues were then refined into rule-based responses. We applied the same process across all scenarios, and experts reviewed a subset of the responses and conversations. We also set the BDI values to reflect the specific settings and goals of the virtual children. Although the BDI values vary slightly across scenarios, we expect the underlying knowledge gained from interaction to be transferable. Participants, therefore, encounter different interactions within the same session to avoid repeating the same inputs.

DELIVERING FEEDBACK

The system provides feedback to guide participants in improving their performance and gaining a better understanding of the situation. We created feedback covering the two aforementioned levels: task-level, which focuses on the current state and actions within the interaction, and process-level, which relates to the broader structure and strategy of the interaction. The feedback is grounded in the five-phase model and is designed to be interpretable, explicitly linking the participant's actions to the child's behaviour.

The system delivers feedback both immediately during the interaction and after the conversation has ended, as combining both approaches has been recommended for effective learning [84, 158]. It addresses the three feedback questions defined by [158] by making expectations explicit (feed up), assessing where the interaction fell short (feedback), and offering suggestions on how to move forward (feed forward). This shifts participants' focus from guessing what to do next toward understanding which strategies are effective and why. We designed the feedback following guidelines on formative feedback [341, 394]. This feedback operates independently of specific scenarios, focusing instead on the applicability of the five-phase model and on how the child's behaviour and thinking change. C.2 includes examples of immediate and delayed feedback taken from the system.

Immediate feedback. During the interaction with the virtual child, participants can request immediate feedback through a set of feedback options (Figure 3.2). These options provide guidance on the child's state, corrective input on the participant's performance, and control over the progress of the conversation. Eight options are available: three at the task level, four at the process level, and one providing general information. Table 3.3 lists these feedback options, their descriptions, levels, and focus areas. The feedback is learner-controlled, allowing participants to request help only when needed. Through this design, we aimed to make information easier to process and more manageable by giving participants agency over when and how to receive support. Prior work suggests that learner-controlled feedback, particularly in terms of timing and performance, is effective in supporting learning outcomes [82, 83]. Furthermore, not all trainees are alike, and they might use feedback differently [71]. This design also allows us to observe when participants choose task- or process-level feedback.

Delayed Feedback. Participants receive delayed feedback after completing an interaction in the guided session. This feedback covers both task and process levels, highlighting potential shortcomings during the conversation (e.g., not showing enough empathy) and their consequences (e.g., the child leaving due to lack of trust). It also provides suggestions for improvement in subsequent interactions, such as asking more clarifying questions.

TRAINEES REFLECTIONS

After completing the interaction and receiving feedback, participants are asked to reflect on the interaction, the feedback, and how these elements shape their overall learning process. The focus is mainly on their reflections about the exercise itself and the learning outcomes of the conversation. We followed Ash and Clayton [22]'s recommendation to write reflective questions that guide learners to consider what they learned, how they learned it, why it matters, and how they can improve.

Table 3.3: Description of the immediate feedback options in the training system and how they guide trainees (i.e., participants) during training.

Feedback Option	Description	Level	Focus
Current Child's Goal	Displays what the child wants to achieve in the conversation (i.e., their desire from the BDI).	Process	Child's thinking
Current Child's Situation	Explains what the child currently thinks or feels based on their beliefs in the BDI.	Task	Child's thinking
Current Phase Progress	Shows the trainee's progress in the current phase (percentage completed) and indicates what is needed to proceed to the next phase.	Process	Trainee's interaction
Advice on Next Message	Provides a hint about what the trainee should type next in the interaction.	Task	Trainee's interaction
View Messages Evaluation	Displays which trainee messages might work better in a different phase. Shows the message, the phase it was used in, and the phase where it would be more appropriate.	Process	Trainee's interaction
5-Phase Model Summary	Provides a summary of the five-phase model, which remains the same throughout the training session.	Overall summary	Static information
Restart the Phase	Resets the entire current phase of the five-phase model. This deletes all messages exchanged during the current phase and restores the child's BDI to its state at the beginning of that phase.	Process	Conversation
Undo My Last Message	Deletes only the most recent trainee message and the corresponding virtual child reply, then restores the child's BDI to the state just before that exchange.	Task	Conversation

A TRAINING SESSION FLOW

A model illustrating the structure of a learning session with the training system is presented on the left side of Figure 3.3. This model is based on Kolb's learning cycle [210] and the Formative Learning Cycle, which integrates feedback, by Brookhart [61] (adapted from [267]). The session begins with a theoretical explanation of the learning objectives. This is followed by an exploration mode, in which participants apply the knowledge gained from the explanation step. The purpose of the exploration mode is to provide a task that resembles real interactions but allows participants to explore without performance assessment, helping them understand and apply the concepts. In the *S* condition, participants received only the simulation and were asked to explore. In the *SF* and *SFR* conditions, they interacted with a simulation that incorporated immediate guiding feedback and received delayed feedback based on their performance during practice. Participants in the three conditions completed the exploration mode twice using the same scenario. In the next step,

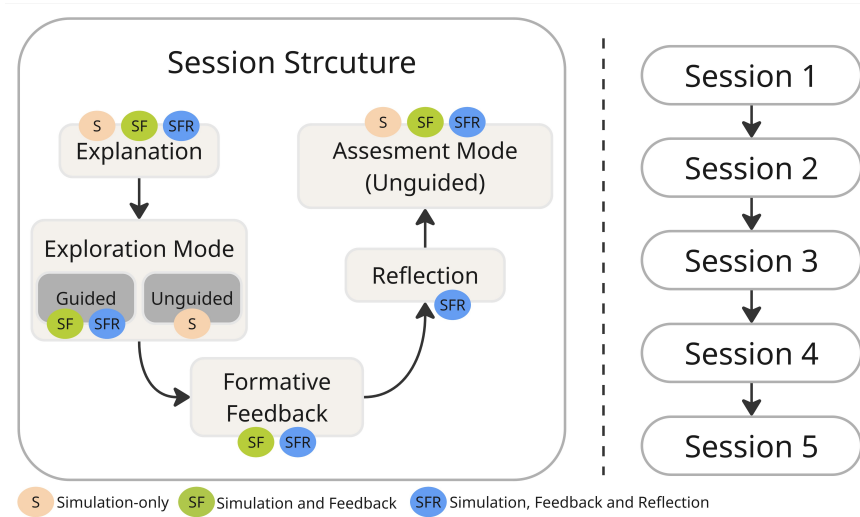


Figure 3.3: Structure and flow of a learning session within the training system, highlighting moments for exploration (exploring the simulation with or without feedback), reflection, and unguided assessment (assessing performance with the simulation only). Each session of the five sessions had the same structure. The design follows Kolb’s experiential learning cycle [210] and Brookhart [61]. Each session maintains the same overall structure, while the specific content varies across the three conditions (S, SF, and SFR).

participants in the *SFR* condition reflected on their learning during the exploration mode. This reflection was meant to help them review and consolidate what they had learned, deepening their understanding of the task. Afterwards, the participants in the three conditions completed the assessment mode, which evaluated their performance as they interacted with the simulation without guidance to apply their knowledge independently.

To bring all components together, Figure 3.4 presents an architecture based on the ARTES architecture [8]. The virtual child’s thinking includes the simulation parts, which control the BDI model and the reasoning process (explained in detail in [145]). The educational model includes the tutor, who provides guiding feedback and determines when to introduce reflection moments. To deliver feedback and guidance, the tutor draws on three sources: the learner’s knowledge component, to assess the participant’s current knowledge and performance; the BDI decision-making, to obtain the virtual child’s state and goals; and the domain knowledge component, to extract relevant feedback from the learning material (e.g., the five-phase model).

3.3.3 MEASURES

We classify the measures as primary and secondary. We used the primary measures to test the hypotheses and address the research questions, and the secondary measures to explore participants’ perceptions of the conditions. All questionnaire items and rubrics are referenced here or provided in C.3.

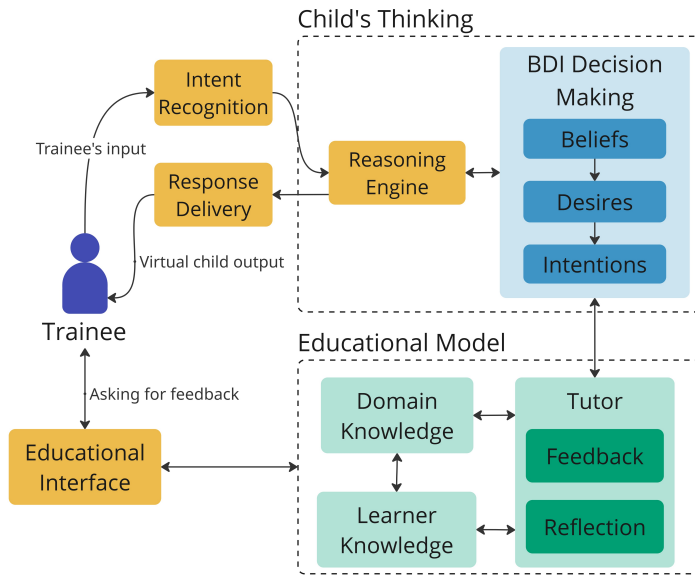


Figure 3.4: System architecture based on the ARTES framework [8]. It integrates the virtual child’s thinking (BDI-based reasoning and response generation) with the educational model, where the tutor component provides feedback and reflection moments informed by learner knowledge, the BDI model, and domain knowledge.

PRIMARY MEASURES

Task Knowledge (a). This measure assessed participants’ knowledge of applying the five-phase model. We developed two versions of a knowledge assessment exam, based on training materials from the Dutch children’s helpline and a handbook on the five-phase model [344]. Each exam version contained different multiple-choice questions, with a maximum score of twelve points. The questions were designed to assess one of four levels in Bloom’s taxonomy [42, 43] (remembering, understanding, applying, and analysing), and this structure was the same across both versions. For each participant, the order of the two exam versions was randomly assigned to serve as pre- and post-tests. After each exam, participants rated their confidence in their answers on a scale from 0 (not confident at all) to 10 (extremely confident), to provide additional insights.

Reflection Capability (b). We used this measure to compare participants’ reflections between conditions at the end of the experiment. Following a similar approach to Moussa-Inaty [270], who categorised participants’ reflections into levels, we assigned reflection scores to participants. To derive these scores, we developed a rubric adapted from the levels of reflection by Fleck and Fitzpatrick [126], Hatton and Smith [159]. Instead of treating the levels as a hierarchy, we used them as five dimensions of reflection: revisiting, explanation, exploring relationships, fundamental changes, and wider implications. Each dimension was scored from 0 to 3, depending on whether no statement was present (0), or one, two, or three or more statements were present in the participant’s written reflection. The maximum total score was 15 points.

Conversational Outcomes (c). The performance measure assessed how successful participants were at applying the five-phase model during their conversations. It represents a score for the participant's interaction with a virtual child in an unguided assessment task, resulting in a separate score for each of the five sessions. To measure the score, we calculated the deviation of the virtual child's belief values from the target belief values at the end of the conversation, that is, what the child's BDI values should be when a trainee applies the five-phase model correctly. In the BDI model, all belief values range from 0 to 1. For binary beliefs (restricted to 0 or 1), a fixed penalty of 0.2 was applied instead of a variation score when the condition to change the belief was not met. All deviations were summed to yield a total score. The optimal deviation score is zero, indicating no divergence from the targeted conversational path. Higher scores indicate greater divergence from the targeted path. Participants were not penalised for achieving better belief values (e.g., showing more empathy than required), as this still demonstrated correct application of the model. This measure is therefore not intended as a direct estimate of real-world counselling quality, but as a proxy for whether trainees can apply the five-phase model to move virtual children toward desirable conversational states. Because the system was implemented across twelve different scenarios, the score reflects application of the protocol rather than performance on a single fixed scenario.

Feedback-Asking Behaviour (RQ1-RQ2). We tracked participants' use of the eight feedback options (Figure 3.2) throughout the experiment. In each guided interaction, we recorded how often each feedback option was selected. This resulted in a count of how many times an option was selected per session. To ensure consistency in the within-subject analysis, we collected this measure only in the condition with all features (simulation, feedback, and reflection).

SECONDARY MEASURES

We examined secondary measures to describe participants' perceived experience. In addition, we collected data outside the scope of this study (e.g., participants' learning strategies, self-efficacy, and perceived values from the virtual child), which we leave for future research (see preregistration for details).

Perceived Usefulness. This measures participants' perception of the training's usefulness. We developed a six-item questionnaire focusing on how useful the information provided by the tool was and how it supported application of the five-phase model. The questionnaire was adapted from the performance expectancy construct in the Unified Theory of Acceptance and Use of Technology model [383] and from the work of Fitrianie et al. [123]. Participants rated each item on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree), with intermediate points at 2 (disagree), 3 (somewhat disagree), 4 (neither agree nor disagree), 5 (somewhat agree), and 6 (agree). An example item is: "The exploration mode provides me with information about the five-phase model".

Overall Experience with the Agent. This qualitative measure assessed how participants perceived their interaction with the agent. We used the 24-item short version of the Artificial Social Agent Questionnaire (ASAQ) [124, 125]. Following the authors' recommendations, responses were rated on a scale from -3 (Disagree), through 0 (Neither agree nor disagree), to +3 (Agree). In all items, we used the term "virtual child" to refer to the agent.

Usability of the Training System. We measured the usability of the training system using the System Usability Scale (SUS) questionnaire [60]. Ratings were made on a 5-point Likert scale: 0 (strongly disagree), 1 (somewhat disagree), 2 (neither agree nor disagree), 3 (somewhat agree), and 4 (strongly agree). In the questionnaire, we referred to the system as “the training system.”

3.3.4 PROCEDURE

Participants were screened to ensure they met the recruitment criteria. Upon confirmation, they were redirected to Qualtrics, an online survey platform, to answer demographic questions and then watch an educational video on the five-phase model, which explained the model’s merits and applications. Afterwards, participants completed an attention check test to verify whether they had watched and understood the video. If a participant failed this test, they had to rewatch the video and retake the attention test. We excluded participants who failed on their second attempt. Following the attention checks, participants completed one version of a knowledge assessment exam, which was randomly assigned, to measure their baseline knowledge of the five-phase model before the experiment.

Participants were then assigned to one of four conditions: no intervention (*C*), simulation only (*S*), simulation with feedback (*SF*), or simulation with feedback and reflection (*SFR*). Participants in the no-intervention condition (condition *C*) did not receive the simulation-based training. For the intervention conditions (*S*, *SF*, and *SFR*), participants received instructions on how to interact with the tool, depending on their assigned condition. For example, participants in the simulation-only (*S*) condition did not receive instructions on the feedback options and what each option does. We also explicitly instructed participants in *S*, *SF*, and *SFR* not to use large language models (LLMs) to interact with the virtual child. After completing comprehension checks, participants began the first training session.

Each session began with an exploration mode and ended with an assessment mode, as explained in Section 3.3.2. The modes were clearly labelled to participants as the “exploration” and “assessment” modes. In the exploration mode, participants interacted with a virtual child either without feedback (condition *S*) or with feedback options in a guided interaction setting (conditions *SF* and *SFR*). They completed two interactions with the same virtual child scenario within this mode. Participants in conditions *SF* and *SFR* received immediate feedback during the interactions and delayed feedback after each interaction. We recorded participants’ use of feedback options during the exploration mode for condition *SFR*. After the exploration, participants in condition *SFR* were asked to reflect on their learning experience. Next, all participants in conditions *S*, *SF*, and *SFR* moved to the assessment mode, where they engaged in an unguided interaction (without feedback options) with a new virtual child scenario, applying the five-phase model. After completing this mode, the interaction session ended. At the end of the experiment session, participants in all four conditions (including *C*) filled out a self-efficacy questionnaire (outside the scope of this study) and had the option to provide further comments on the session.

After completing a session, participants were invited to the next session, reassigned to the same condition as before, and followed the same procedure. We invited participants in condition *C* to complete post-questionnaires again after twelve days, with 48 hours to complete the session (i.e., fourteen days in total). In the other three conditions, each participant completed five sessions over fourteen days. For each session, participants had

up to 48 hours to begin after receiving the invitation and were required to complete it in a single sitting without a timeout. After the 48-hour window, there was a 24-hour break before participants were invited to the next session.

Within the five sessions, we randomised two factors to reduce potential bias and counterbalance order effects. First, participants received 10 out of 12 virtual child scenarios in a random order, such that each scenario was shown only once (two scenarios per session: one in the exploration mode and one in the assessment mode). Second, for participants in the feedback conditions (*SF* and *SFR*), we randomised the order position of the feedback option buttons (Figure 3.2) between participants, but kept the order identical for each participant across all five sessions. To ensure coherence, relevant options were paired together according to their focus (Table 3.3; e.g., "Undo My Last Message" and "Restart the Phase").

Upon finishing the fifth session, participants completed the post-questionnaires. Here, they first took the alternate version of the knowledge assessment exam as a post-exam. They were then asked to reflect on their overall learning experience and complete the secondary measures questionnaires. Before concluding the experiment, participants were asked two quality control measures: whether they had used an LLM during the experiment (and if so, why), and whether they recommended that their data be used for scientific purposes. Figure 3.1 provides an overview of the experiment's procedure and the measures used at each stage.

3.3.5 DATA ANALYSIS

For quality control measures, all participants recommended using their data for scientific purposes. Regarding the question of whether they used LLMs, twelve participants reported using LLMs to help in brainstorming during the chat ($n = 2$), writing reflections ($n = 6$), getting information about the five-phase model ($n = 2$), answering the knowledge test ($n = 1$), and translation ($n = 1$). We included their data for analysis, as they did not use LLMs to generate responses to the virtual child.

To obtain a final score for the questionnaire measures, we calculated the mean scores per scale, reverse-coded items where appropriate, and followed the guidelines provided for the questionnaires. The conversational outcomes score reflects deviation from the optimal path, where higher scores indicate worse performance. For easier interpretation and analysis, we reversed the score and transformed the data into a performance index using a linear transformation based on the possible score range, i.e., $(\text{max score} - \text{observed score}) / \text{score range} \times 100$. This index is presented on a percentage scale: 0 represents no performance improvement, i.e., the child's starting position, and 100 represents the best performance, correctly applying all learning goals. For the SUS questionnaire, one item was not collected due to an administrative error ("*I thought the system was easy to use*"). We corrected this by multiplying the resulting score of the other nine items by 2.78 (instead of 2.5) to obtain a final score out of 100. For task knowledge and reflection, item scores were summed to form total scores, ranging from 0 to 12 and from 0 to 15, respectively.

We used a local LLM (LLaMA 3.1 8B Instruct model) to evaluate participants' reflections, given the large number of reflections (346 reflections). Our approach was inspired by Chiang and Lee [78], who tested LLMs to score open-ended story texts using dimensions

such as likeability, relevance, and cohesiveness. They found the results to be consistent between human experts and LLM evaluations. We also followed Bavaresco et al. [31]’s recommendation to check the reliability of LLM evaluations against human evaluators. To do this, we prompted the LLM with the scoring rubric, explanations of the rubric dimensions, and five example reflections that included human-assigned scores and justifications. After receiving this prompt, both the LLM and the first author independently scored a different set of 100 reflections to assess consistency between human and model ratings. The resulting LLM scores showed a correlation of $r = 0.80$ with the human-coded final reflection scores, which is classified as a high correlation by Hinkle et al. [165]. To further check for individual biases, a second coder (a PhD candidate in computer science) independently scored the same 100 reflections after training on the five examples that were given to the LLM. Agreement between the first and second human coders was 0.77 (high correlation), while agreement between the LLM and the second coder was 0.69 (moderate correlation), following the same classification [165]. Given this reliability, we used the LLM to score all the reflections in our analysis.

First, we checked whether people improved over time. We analysed the within-condition changes in task knowledge (pre- vs. post-test) and conversational outcomes (first session vs. last session) using Bayesian paired-samples t -tests. As for hypothesis testing between conditions, we also analysed the data using Bayesian methods. To predict the task knowledge post-test, we fit a multilevel model with fixed effects for condition and pre-test scores, and random intercepts for participants. For conversational outcomes collected across multiple sessions, we tested the added value of each condition per session. Thus, we fitted a multilevel model with fixed effects for session, condition, and their interaction, and random intercepts for participants to account for the repeated measures. To compare the overall increase in conversational outcomes between conditions, we computed the delta per participant (i.e., the difference between the indexes of the first and last sessions). Then we fitted a model to assess how the conditions predict the index delta for the conditions. For reflection and other post-only measures (e.g., perceived usefulness), we fit linear models with condition as a fixed effect. To test for an overall increase with the subsequent addition of elements (H_4), we fitted the same models using an ordered categorical predictor with a monotonic effect for conditions. Regarding the two explorative research questions, we examined trends in feedback options and performance over time, as well as the associations between feedback options usage and learning outcomes.

For all Bayesian analyses, we used uninformative priors and interpreted posterior probabilities to evaluate hypotheses, following the guidelines of Andraszewicz et al. [16], Chechile [76]. We report 95% Highest Density Intervals (HDIs) to indicate the range of credible parameter values. We inferred effect size probabilities based on the Sequential Effect eXistence and sIgnificance Testing (SEXIT) framework [241]. Following the framework, we report the proportion of the posterior distribution lying beyond predefined thresholds for “small” and “large” effects (set at $|0.10|$ and $|0.30|$ of the outcome’s standard deviation, respectively) [132, 241]. We conducted all analyses in R (version 4.3.1) using the `brms` package [67] for multilevel modelling, the `bayestestR` package [242] to calculate HDIs and effect sizes, and the `BayesFactor` package [266] for Bayesian t -tests. The data, R code for the analysis, generated markdown reports, Docker image for reproducibility, and

prompts are publicly available via the 4TU Research Data Repository². The system code is also publicly available on GitHub³.

3.4 RESULTS

3.4.1 HYPOTHESES RESULTS

Before testing the hypotheses, we examined whether the four conditions showed improvement over time. Table 3.4 presents the results for the two primary repeated measures (task knowledge and conversational outcomes) within each condition. The table also shows means and standard deviations for all three measures in each condition. A summary of the four hypothesis results is presented in Table 3.5, including the probability that each hypothesis is true, its interpretation, and the likelihood that the observed effect sizes are small or large.⁴

Within-condition results: Posterior probabilities in Table 3.4 show credible support that participants in the base control condition (*C*) decreased in task knowledge over time, with a 30% probability of a large effect size. The condition *S*, which involved only the simulation, showed some support for improvement: there was a 91% posterior probability that task knowledge increased over time, with a 13% probability of a large effect size, and a 92% posterior probability for an increase in conversational outcomes over time, with a 15% probability of a large effect size. For participants in conditions *SF* and *SFR*, the results show credible support for improvements over time in both task knowledge and conversational outcomes. For *SF*, the probabilities of a large effect size are 89% and 99.45%, respectively, for task knowledge and conversational outcomes. For *SFR*, the probabilities of a large effect size are 37% for task knowledge and 77% for conversational outcomes.

H_4 : Overall: subsequent incorporation leads to higher measures: We begin with this hypothesis to test whether there is an overall effect of sequentially adding components. Across all measures, the results showed credible support for the subsequent incorporation of components (Table 3.5). The corresponding posterior probabilities (task knowledge: > 99.99%, reflection capability: 99.93%, and conversational outcomes: 96%) provide evidence for overall positive effects. Hence, H_4 is supported. Regarding the likelihood of a large effect size, it is almost certain for task knowledge (99%), somewhat lower for reflection capability (87%), and lower for conversational outcomes (35%), for which a small effect appears more likely (81%).

H_1 : *S* is higher than *C*: Participants in the simulation-only condition showed higher improvement than those in the control condition, both in task knowledge and reflection capability (Table 3.4). As shown in Table 3.5, the posterior probabilities (99.9% and 98.9%) provide credible support that the *S* condition improved both measures more than the *C* condition, supporting H_1 . The likelihood of large effect sizes also supports this outcome, with 92% for task knowledge and 70% for reflection capability.

H_2 : *SF* is higher than *S*: Table 3.5 shows 89% posterior probability that participants in the simulation with feedback condition (*SF*) had higher task knowledge than the simulation only (*S*) condition. Reflection capability was similar between the two conditions, with a

²<https://data.4tu.nl/datasets/635d08b5-814e-45d3-99c1-c330f3c9813d>

³<https://github.com/alowayyedm/LilobotTraining/tree/BDIFeedbackFinal>

⁴All model comparison results and additional analyses are provided in Table C.5 in the appendices.

Table 3.4: Descriptive statistics and Bayesian within-condition test results for task knowledge (pre- vs. post-test) and conversational outcomes (session 1 vs. session 5) across experimental conditions.

Condition	Measure	% Posterior Probability > 0	Interpretation	Estimate [95% HDI] (SD)	% Effect Size >	
					Small	Large
C	Task Knowledge M _{pre} = 6.41 (1.66) M _{post} = 5.92 (1.84)	0.03	Good bet against- too good to disregard	-0.48 [-0.98, 0.01] (0.25)	0.86	0.30
	Reflection M _{post} = 3.67 (3.01)	-	-	-	-	-
	Conversations	0.91	A promising but risky bet	0.36 [-0.18, 0.88] (0.27)	0.70	0.13
S	Task Knowledge M _{pre} = 6.70 (1.55) M _{post} = 7.06 (1.95)	0.92	A promising but risky bet	4.34 [-1.79, 10.34] (3.09)	0.73	0.15
	Conversations M _{s1} = 41.09 (20.04) M _{s5} = 45.50 (22.03)	-	-	-	-	-
	Reflection M _{post} = 4.97 (3.61)	-	-	-	-	-
SF	Task Knowledge M _{pre} = 6.38 (1.73) M _{post} = 7.38 (1.90)	0.9998	Nearing certainty	0.99 [0.46, 1.52] (0.27)	>0.99	0.89
	Conversations M _{s1} = 46.64 (20.57) M _{s5} = 64.13 (23.87)	>0.99995	Virtually certain	17.26 [10.56, 24.10] (3.45)	>0.99	>0.99
	Reflection M _{post} = 4.97 (2.80)	-	-	-	-	-
SFR	Task Knowledge M _{pre} = 6.71 (1.73) M _{post} = 7.39 (2.00)	0.9996	Nearing certainty	0.68 [0.28, 1.07] (0.2)	0.98	0.37
	Conversations M _{s1} = 46.64 (22.95) M _{s5} = 58.56 (26.70)	>0.99995	Virtually certain	11.85 [6.58, 17.07] (2.68)	>0.99	0.77
	Reflection M _{post} = 5.43 (3.50)	-	-	-	-	-

Note. C = Control; S = Simulation only; SF = Simulation and Feedback; SFR = Simulation, Feedback and Reflection. Reflection capability was measured at post-test only; therefore, no within-condition change statistics are reported.

Table 3.5: Summary of Bayesian test results for the four hypotheses, comparing conditions across the three outcome measures: task knowledge, reflection capability, and conversational outcomes.

Hypothesis	Measure	% Posterior Probability > 0	Interpretation	Estimate [95% HDI] (SD)		% Effect Size >	
				Small	Large	Small	Large
H₁: Overall Increase	a: Task Knowledge	>0.99995	Virtually certain	1.38 [0.82, 1.92] (0.28)	>0.99	>0.99	>0.99
	b: Reflection Capability	0.99935	Very strong bet	1.56 [0.62, 2.52] (0.49)	0.99	0.99	0.87
	c: Conversations Overall	0.93	A promising but risky bet	7.12 [-2.34, 16.13] (4.73)	0.81	0.81	0.35
	c: Conversations per Session	0.96	Good bet for - too good to disregard	1.22 [-0.14, 2.24] (0.60)	0.01	0.01	< 0.01
H₁: S > C	a: Task Knowledge	0.9992	Very strong bet	1.05 [0.41, 1.70] (0.33)	>0.99	>0.99	0.92
	b: Reflection Capability	0.989	Good bet for - too good to disregard	1.31 [0.18, 2.44] (0.58)	0.95	0.95	0.70
H₂: SF > S	a: Task Knowledge	0.89	Only a casual bet	0.41 [-0.26, 1.05] (0.33)	0.72	0.72	0.26
	b: Reflection Capability	0.5	Not worth betting	-0.006 [-1.14, 1.10] (0.58)	0.28	0.28	0.04
	c: Conversations Overall	0.9937	Strong bet	13.11 [2.85, 23.30] (5.22)	0.98	0.98	0.78
	c: Conversations per Session	0.9954	Very strong bet	2.99 [0.72, 5.25] (1.16)	0.68	0.68	< 0.01
H₃: SFR > SF	a: Task Knowledge	0.40	Not worth betting	-0.07 [-0.63, 0.47] (0.28)	0.32	0.32	0.03
	b: Reflection Capability	0.83	Only a casual bet	0.47 [-0.49, 1.44] (0.49)	0.60	0.60	0.14
	c: Conversations Overall	0.11	Only a casual bet against	-5.54 [-14.00, 3.68] (4.50)	0.71	0.71	0.22
	c: Conversations per Session	0.14	Only a casual bet against	-1.06 [-2.92, 0.92] (0.98)	0.08	0.08	< 0.01

Note. C = Control; S = Simulation only; SF = Simulation and Feedback; SFR = Simulation, Feedback and Reflection.

posterior probability of 50%. The conversational outcomes showed credible support that the *SF* condition improved more across sessions than the *S* condition, with a 99.5% posterior probability and an estimated average difference of 2.99 per session (95% HDI [0.72, 5.25]). This was further supported by a 78% likelihood that the overall increase in conversational outcomes reflects a large effect size. The pattern is also illustrated in Figure 3.5, where the average indexes per session for the *SF* condition remain consistently better than those for *S*. Overall, these results provide credible support for an effect of adding feedback to simulation on conversational outcomes, but only some support for additional improvements in task knowledge, supporting H_2 .

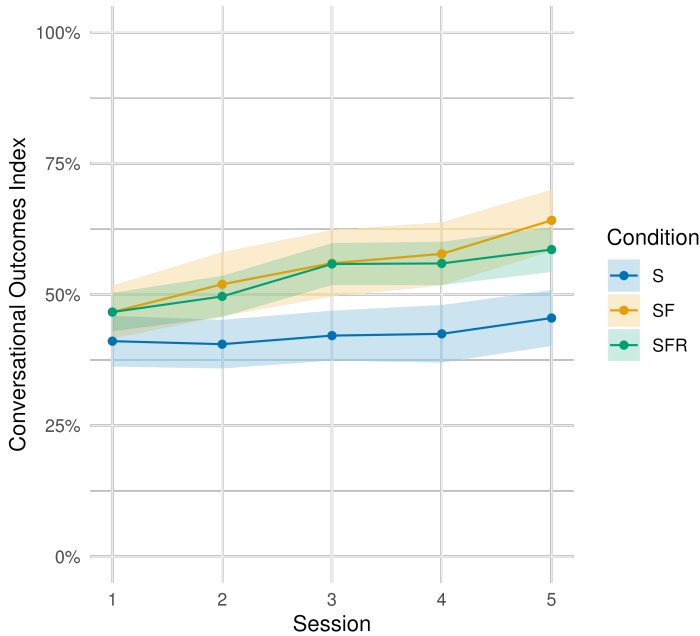


Figure 3.5: Average conversational outcome deviation indexes per condition across sessions. The measure was collected during each Assessment Mode interaction, yielding one index per participant (see Figure 3.3). Shaded areas represent 95% credible intervals.

H_3 : SFR is higher than SF: As shown in Table 3.5, participants in the simulation with feedback and reflection condition (*SFR*) performed differently from those in the simulation with feedback condition (*SF*) in some measures. For task knowledge, the posterior probability is 40%, interpreted as not worth betting on. Reflection capability showed some support for improvement in the *SFR* condition, with an 83% posterior probability and a 14% likelihood of the effect size being large. In contrast, conversational outcomes provided some support against the *SFR* condition, with a 14% posterior probability (only a casual bet against; 86% support for the opposite) that *SFR* increases conversational outcomes compared to *SF*. This pattern is also visible in Figure 3.5, where participants in *SF* scored on average higher than those in *SFR* in the final session. Taken together, these findings suggest that adding reflection to feedback improved reflection capability (some support for H_3c),

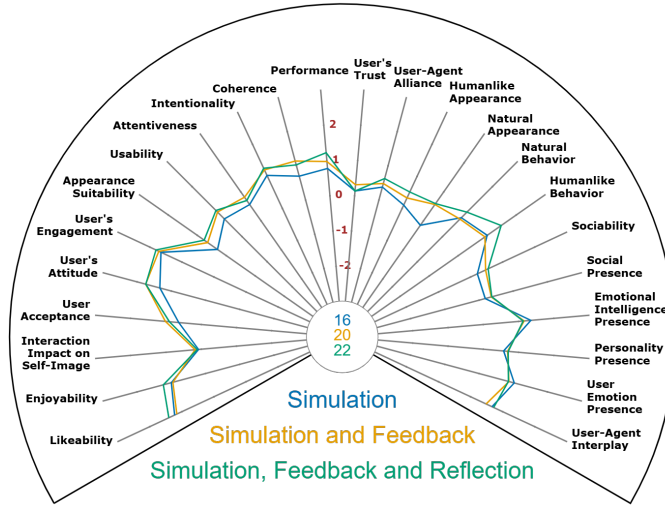
but led to a reduction in conversational outcomes improvements (some support against H3b), and no additional task knowledge improvement (no support for H3a). Therefore, H_3 as a whole does not seem to hold.

3.4.2 SECONDARY RESULTS

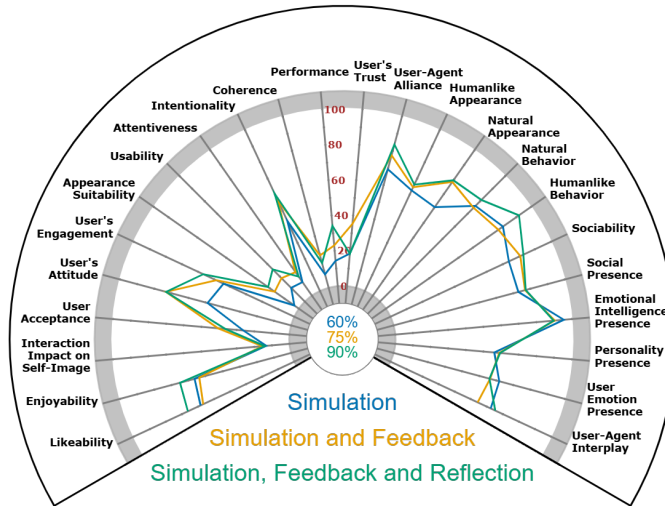
For the secondary analysis, we asked participants in the conditions *S*, *SF*, and *SFR* to rate their perceived usefulness of the training tool, their overall perceived experience with the agents (through the short version of ASAQ), and the usability (SUS) of the system. Participants perceived the usefulness of the three conditions differently (*SFR*: $M = 6.16$, $SD = 0.78$; *SF*: $M = 6.12$, $SD = 0.83$; *S*: $M = 5.36$, $SD = 1.01$). The analysis shows credible support that adding feedback increases perceived usefulness (posterior probability: $> 99.99\%$, 95% HDI [0.46, 1.04]), but no support for adding reflection to feedback (posterior probability: 66%, 95% HDI [-0.2, 0.29]). Next, the overall experience ratings (ASAQ) also differed (*SFR*: $M = 0.91$, $SD = 1.2$; *SF*: $M = 0.83$, $SD = 1.21$; *S*: $M = 0.68$, $SD = 1.03$), with some support for adding feedback (posterior probability: $> 78\%$, 95% HDI [-0.24, 0.56]), and no support for adding reflection (posterior probability: $> 68\%$, 95% HDI [-0.25, 0.44]). Figure 3.6a shows how the three conditions compare across each construct of ASAQ, and Figure 3.6b compares the three conditions to the 29 artificial social agents in the ASAQ Representative Set 2024 [125]. The *S* condition scored with overall ASAQ higher than 60% of the dataset, while *SF* scored 75%, and *SFR* 90%. Regarding SUS, participants scores also differs (*SFR*: $M = 78.33$, $SD = 14.33$; *SF*: $M = 77.08$, $SD = 14.54$; *S*: $M = 73.19$, $SD = 15.44$). All three scores fall within the “good” usability range [26]. The statistical analysis on SUS scores shows that adding feedback increased usability (posterior probability: $> 93\%$, 95% HDI [-1.06, 8.97]), and there is some weak support for adding reflection compared to simulation and feedback (posterior probability: $> 72\%$, 95% HDI [-3.00, 5.54]).

3.4.3 EXPLORATORY RESULTS

RQ1: Using feedback levels over time. To examine whether participants used task- or process-level feedback options differently over time, we analysed their usage patterns across sessions in condition *SFR*. The results did not indicate a clear tendency toward either type (posterior probability = 0.72, $M = 0.92$, 95% HDI = [-2.21, 4.06]). Therefore, we visualised the data to explore potential trends across specific feedback options and participants (Figure 3.7). As shown in the figures, participants generally used both feedback types less frequently over time, which is consistent with the analysis of feedback option requests over time (posterior probability = 0.99, $M = -2.20$, 95% HDI = [-4.02, -0.36]). Figure 3.7a shows that the most frequently used feedback options were *Advice on next message* (task level) and *Current phase progress* (process level), whereas options that directly affected the conversation—*Undo last message* (task level) and *Restart phase* (process level)—were used the least. In Figure 3.7b, we categorise each interaction in the exploration mode (one per participant per session) into one of five feedback options categories based on the dominance of task- or process-level feedback: process-dominant ($\geq 25\%$ process-level use), slightly process-leaning (10–25% process level), balanced (within $\pm 10\%$), slightly task-leaning (10–25% task level), and task-dominant ($\geq 25\%$ task level). In the first session, some participants used more task- or process-level feedback, but the largest proportion (40%) fell into the balanced category. Over time, the proportion of participants with balanced



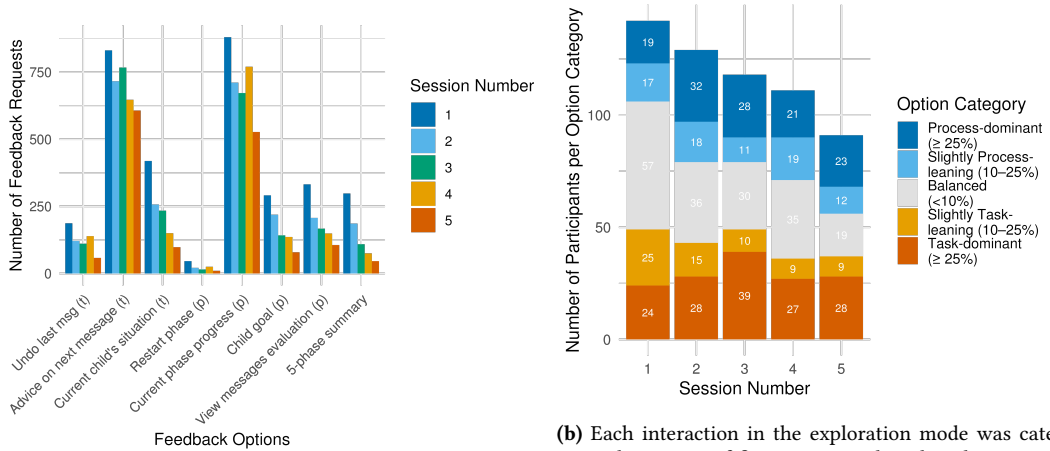
(a) Comparison of the raw ratings for the short ASAQ between the three conditions.



(b) Percentile chart of each condition on the short ASAQ constructs, normalised against the 29 agents in the ASAQ Representative Set 2024 [125]. Grey bands indicate values lower than or higher than those observed for all agents in the reference set.

Figure 3.6: Two comparison views based on the short ASAQ constructs. The charts were generated using the ASAQ visualisation tool.

feedback requests decreased to 21% in session five, as participants increasingly shifted their feedback requests toward either task- or process-level categories.



(a) Number of times each feedback option was clicked per session. (b) Each interaction in the exploration mode was categorised into one of five categories based on how many times a participant clicked task- or process-level feedback.

Figure 3.7: Overall description of feedback option usage across the five sessions.

RQ2: Influence of feedback level on knowledge, reflection capability and conversational outcomes.

We tested whether using more task-level feedback than process-level feedback (or vice versa) was associated with better outcome measures. No apparent association effects were found for any of the three measures: task knowledge, reflection capability, and conversational outcomes. Specifically, when comparing the relative use of each level, the posterior distributions showed weak tendencies toward more task-level feedback for knowledge (74%, $M = 0.64$, 95% HDI = [-1.27, 2.55]), but a weak tendency toward more process-level feedback for reflection (61%, $M = -0.38$, 95% HDI = [-3.16, 2.32]), and conversational outcomes over the five sessions as an interaction effect (68%, $M = -1.21$, 95% HDI = [-6.08, 3.72]). All three distributions are interpreted as not worth betting on. However, when examining only the main effect on conversational outcome indexes (without the interaction with session), there is some support that using task-level feedback led to better conversational outcomes compared to process-level (93%: a promising but risky bet, $M = 12.52$, 95% HDI = [-3.96, 29.17]). This effect appeared most prominently in the first session. The overall patterns for these three measures are shown in Figure C.4 in the appendices, which illustrates how the feedback option categories scored across the three outcome measures.

3.5 DISCUSSION AND CONCLUSIONS

In this paper, we examined how the subsequent addition of simulation (H_1), feedback (H_2), and reflection (H_3) shapes trainees' learning in simulation-based training. Across five sessions, we assessed how these elements contributed to improvements in task knowledge,

reflection capability, and conversational outcomes (performance). Overall, the results support the cumulative effect proposed in H_4 : adding elements sequentially leads to overall increases across the three learning measures. Still, each element contributed differently, and these differences highlight important considerations for the design and impact of simulation-based training.

Participants interacting with the simulation only (S) showed increases over time in both task knowledge and conversational outcomes, whereas the control group (C) showed a decrease in task knowledge. H_1 results indicate that adding simulation increases knowledge and reflection more than the control condition. This improvement may be attributed to the effect of repetition in several sessions, which supports learning [386]. Participants in condition S needed to repeatedly decide what to ask, how to phrase questions, and how to move the conversation forward across five sessions, whereas the control condition received only theoretical explanations. This lack of repeated engagement may also explain the observed decrease in task knowledge in condition C . The results of condition S align with Reger et al. [313], who found that training with a virtual patient for motivational interviewing—compared to academic study alone—led to increased skill outcomes.

Adding feedback to the simulation (condition SF) improved task knowledge and conversational outcomes, while maintaining similar reflection capability scores compared to the simulation-only condition (H_2). The need for feedback was further highlighted by participant comments, with several participants in condition S explicitly commenting at the end of the sessions that they wished for more guidance, e.g., “I wish there was a redo button...” and “It would be good to have feedback as well”. On the other hand, the results for reflection capability were similar for the two conditions. One possible explanation is that participants across both S and SF received similar amounts of interaction experience with the tool across five sessions, which may have influenced their reflections, even without targeted reflection training. For conversational outcomes, participants across conditions did not start from similar baselines after the first session. Participants in the S condition performed worse in the first session compared to those in SF (Figure 3.5). This is likely because the two conditions with feedback benefited from guidance that helped them navigate their conversations more effectively from the beginning. In turn, this inflated the standard deviation used to compute the thresholds for small and large effect sizes, which resulted in a lower probability of observing effects when comparing improvements in conversational outcomes between S and SF across the five sessions. These baseline differences should be considered when interpreting the conversational outcome results across conditions.

Although adding reflection moments (condition SFR) improved reflection capability as expected, it dampened conversational outcomes compared to simulation with feedback (H_3). The latter may be attributed to two factors. First, the reflection task introduced a break between exploration and assessment modes. This temporal shift in focus, from applying the task in the exploration to reflecting on it, may have hindered immediate carry-over of learning into the assessment interaction — the SF group moved directly from exploration to assessment, while the SFR group completed a reflection task in between. Second, reflection requires effort, motivation, and practice [18, 157, 265]. Yet despite the effort required, many participants in the S and SF conditions were intrinsically motivated to write reflective comments in the optional comments field after each session, even without being prompted. Thus, reflection may have occurred spontaneously for a subset of participants in S and SF .

For the secondary measures, participants reported better perceived usefulness and overall experience as more elements were added. This suggests that trainees valued not only the learning content but also the added structure, guidance, and support within the system. The exploratory research questions on how participants used task- and process-level feedback did not reveal clear patterns; thus, a plausible explanation is that both types should be used depending on the situation and what information trainees need to get. This interpretation is also supported by recommendations to combine the levels of feedback [158], and that the level should focus on the current problem being solved [349]. Task-level feedback, however, was associated with better conversational outcomes when examining individual sessions. This may be due to the “Advice on next message” option, as it directly suggests what to consider next, which can help trainees progress further through the conversation, thereby making the assessment interaction easier to navigate. Future work could examine whether such differences persist days or weeks after the training, as directive guidance might not have long-term benefits for knowledge transfer [320].

3.5.1 LIMITATIONS AND FUTURE WORK

Beyond the five-phase model used in this study, the incremental benefits of simulation, feedback, and reflection are not tied to this protocol alone. They may also extend to training other communication frameworks, such as motivational interviewing [163]. Likewise, the approach may generalise to other helpline and training contexts beyond children’s helplines, such as suicide prevention or mental-health crisis lines, where volunteers must learn to navigate emotionally complex and sensitive conversations. Although the scenarios in this study focused on supporting a child to find a solution, the underlying mechanisms of simulation, guided practice, and supported reflection seem relevant across domains involving interpersonal interactions, e.g., sales, negotiations, or customer service.

The study relied on crowdsourced participants, and more motivated participants (either extrinsically through compensation or intrinsically due to interest) may have been naturally retained across sessions. This could affect the generalisability for trainees who are not motivated. However, in the case of children’s helplines, trainees are typically volunteers who are themselves motivated to learn, which makes this characteristic of the sample less problematic. Additionally, several participants reported misunderstandings by the virtual child. We believe these stem from the rule-based intent recognition component, which did not perform well with verbose input. This limitation might be addressed by integrating Large Language Models (LLMs), as Al Owayyed et al. [7] showed that an LLM-integrated virtual child was perceived more positively than its rule-based counterpart.

Another limitation to generalisability is that excluded participants were slightly more likely to be parents than included participants, although the effect size of this difference was small. This may reflect the five-session design, which required commitment and availability and may therefore have been harder to sustain for participants with children. At the same time, most child helpline counsellors are volunteers who also need to commit time, which may make the retained sample more similar to the target group. Still, this does not threaten internal validity, as parental status did not differ across the four included conditions.

While repeated practice improves learning and retention [193], this study does not address transfer of skills to real helpline communications. In particular, the conversational outcomes measure should be interpreted as a proximal, protocol-aligned performance

measure rather than as direct evidence of transfer to real help-seekers. Thus, results might differ if assessed weeks or months later or in settings involving real help-seekers, especially since reflection supports knowledge retention [73, 219]. Future research could compare simulation-based training with traditional training methods and evaluate knowledge transfer in actual helpline contexts.

Participants in the *SFR* condition showed higher reflection capability. However, it remains unclear whether this improvement reflects growth in counselling-specific reflections or simply a general increase in reflective writing skills. It is possible that participants were practising reflection across the five sessions without necessarily developing deeper counselling insight, as practice improves reflection ability [310]. Future work could test this by adding a condition in which participants reflect on an unrelated topic, such as cycling or daily routines, to differentiate domain-specific reflective growth from general reflective writing skills. Furthermore, since the experiment design was not fully factorial, the independent contribution of reflection (without simulation or feedback) remains unknown and could be examined in future work.

3.5.2 ETHICAL CONSIDERATIONS

Children exhibit diverse behaviours. Modelling only limited behaviours and scenarios in simulations (e.g., only bullying) could risk reinforcing stereotypes about how children act when contacting helplines [8]. Therefore, training should include a wider range of perspectives, behavioural patterns, and helpline topics (e.g., sexuality, family issues, relationships [131]) to avoid narrowing trainees' expectations. For example, it is uncommon for children to ask helplines to call their school on their behalf. Yet, repeated exposure to such a scenario might give trainees the false impression that this is typical.

3.5.3 CONCLUSION

This study shows that simulation-based training can support the skill development of trainee helpline counsellors. Adding feedback enhances the learning process further by helping trainees better understand the child's perspective and their own decision-making within the conversation. Furthermore, adding reflective practice to simulation and feedback strengthens reflection capability, although it appears to slow the rate of performance improvement compared to simulation and feedback alone. When combined, simulation and feedback create a training environment that enhances participants' learning, performance, and experience with the training tool, while incorporating reflection requires more careful consideration. Based on these findings, we hope that simulation-based training, particularly when combined with feedback, can serve as a scalable addition to helpline education, accelerating early-stage counsellor development.

4

TRAINING CHILD HELPLINE COUNSELLORS WITH VALUE-INTEGRATED CHAT SIMULATIONS

4

Children’s helplines train new counsellors 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 behaviour is guided by underlying values. The trainees’ task is to apply motivational interviewing to recognise 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 counsellors for value-sensitive conversations.

Although the published article uses American English, this thesis chapter is presented in British English to ensure consistency with the language style throughout the thesis.

📖 **Al Owayyed, M.**, Brinkman, W.P., Guan, K. W., Keijsers, L., and Tielman, M. L. Training Child Helpline Counselors with Value-Integrated Chat Simulations. *Accepted for publication in International Journal of Human-Computer Interaction.*

4.1 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 [176], training new counsellor 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 [97]. Training typically involves role-playing, with a counsellor 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 [146, 169, 381].

To make such training simulations effective, it is crucial to accurately model the decision-making cues children use during helpline conversations. Human behaviour is influenced by many factors, among which are values [109]. Values guide behaviour explicitly and implicitly [196, 197], shaping how individuals approach interactions and make choices. Like adults, children also hold values that inform their decisions [66, 105], such as a preference for "help others" or "be adventurous". In this paper, we propose enhancing virtual children's behaviour 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, recognising and aligning with these values can lead to better counselling. Helpseekers' values are key in shared decision-making between healthcare providers and those seeking help [46]. Counselling textbooks, both general [251, 292] and those focused on child helplines [344], emphasise the importance of recognising and aligning with the values of helpseekers. Several prominent therapeutic approaches also highlight the role of values. For instance, client-centred therapy prioritises the client's perspective and personal understanding to guide the therapeutic process [317, 318]. Values clarification, meanwhile, helps clients identify, articulate, and actualise their values to guide behaviour and decision-making [203]. Motivational Interviewing (MI) can be seen as a hybrid of client-centred therapy and values clarification [203]. In MI, the counsellor works collaboratively with the client to explore goals and decide on appropriate actions [261]. This method has also been used with adolescents [276], and in helplines [300]. MI incorporates adolescents' perspectives to emphasise their autonomy, which is a key aspect of their development [98].

Virtual agents have employed MI—either to train counsellors through interaction with a virtual patient (e.g., [10]), or to provide counselling via a virtual counsellor (e.g., [354]). Furthermore, a recent randomised trial found that trainees with limited MI knowledge benefited more from virtual patient simulations than from academic study alone [407]. 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 behaviours by applying MI.

For people to recognise 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 [329]. Moreover, incorporating values could enable counsellors

to practice aligning with them and to understand better how values influence a child's decisions. A necessary first step is learning to recognise which values a child might exhibit. This is similar to recognising emotions, which can lead to better adjustment [221, 235]. 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 behaviour can be vague [305], and training with a simulation alone might lack sufficient guidance [8]. Trainees may require explicit guidance through feedback to recognise how a child's values influence their decisions. Providing feedback on a virtual agent's decision-making can help interpret its behaviour [151], leading to a more guided learning experience [204]. 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 counsellors [6, 145]. Lilobot simulates a virtual child contacting a helpline after being bullied at school. The trainee counsellor's task is to follow the five-phase model, a communication protocol commonly used by helplines [344]. Lilobot is built using a BDI (Belief-Desire-Intention) model [137]. 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 counsellors to recognise, understand, and align with children's values. By explicitly modelling 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 counselling skill acquisition or therapeutic effectiveness.

We therefore examine four constructs. First, recognising 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 behaviour provides clear cues that make value-based alignment possible and actionable. This notion relates to the formation of a therapeutic alliance [47, 172] and builds on MI, where counsellors 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 [222, 259]. 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 [383]. 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 (under-

standing 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 modelling, (2) a model with integrated values, and (3) a value-based model accompanied by explanations of the virtual child's reasoning.

4.2 RELATED WORK

4.2.1 AGENTS FOR TRAINING HELPLINE COUNSELLORS

Limited research exists on simulation-based training for helpline counsellors, particularly those for children. In the context of suicide helpline training, Demasi et al. [99] 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) [382]. This game uses a choice-based interaction model to train counsellors in navigating sensitive situations. However, the choice-based design does not allow trainees to practice constructing their own responses, unlike open-ended formats.

4.2.2 MOTIVATIONAL INTERVIEWING AGENTS

Interactive agents have been used to simulate counsellors applying MI. In their scoping review, Mercado et al. [258] identified several cases where conversational agents employed MI to support healthy behaviours, such as addressing alcohol misuse [323], promoting physical activity [291], and improving women's preconception health [178]. Agents delivering MI were also shown to improve conversational skills when used before group discussions [327].

On the other hand, interactive agents have also been used to simulate help-seekers, allowing healthcare providers and counsellors to train in delivering MI. Results from such training systems are promising. For example, Albright et al. [10] evaluated a virtual patient training system for substance abuse counselling, finding increased knowledge and skills among trainees compared to a control group. Similarly, Oster et al. [293, 294] reported that MI training with a virtual client for diet and weight loss counselling improved counsellors' skills. Other applications include training for HIV care [322] and clients hesitant to speak with a therapist [313].

4.2.3 VALUES IN SOCIAL AGENTS

Human values have been modelled in social simulations, which typically involve simulating interactions among multiple agents to study their interactions. For example, Heidari et al. [161] 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 behaviour can form. Other works also integrate values into multi-agent systems (e.g., [89, 378]).

In our case, we simulate values in a single agent interacting with a human user, rather than modelling society-level dynamics. Some frameworks, such as the one proposed by Cranefield et al. [93], 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 [49, 51], the agent's behaviour could be driven by values such as autonomy, respect, or dominance. Similarly, in a vaccination training system [274], 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 [315], 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 [256], for example, a strategic agent might prioritise achievement, while a free agent might value building trust.

4.2.4 CHILDREN'S VALUES

Children's values can affect their decisions and behaviours. 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 [373, 398]. Even at young ages (e.g., 4–8 years old), children can express value priorities such as benevolence, achievement, security, and autonomy [333, 398]. These value priority categories follow the Schwartz theory of basic human values [336], which organises values into higher-order dimensions. This values model has been applied in developmental research to ground children's value orientations (e.g., [66, 105, 333]). 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 [333]. 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 behaviour can support counsellor training through simulations.

4.3 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.

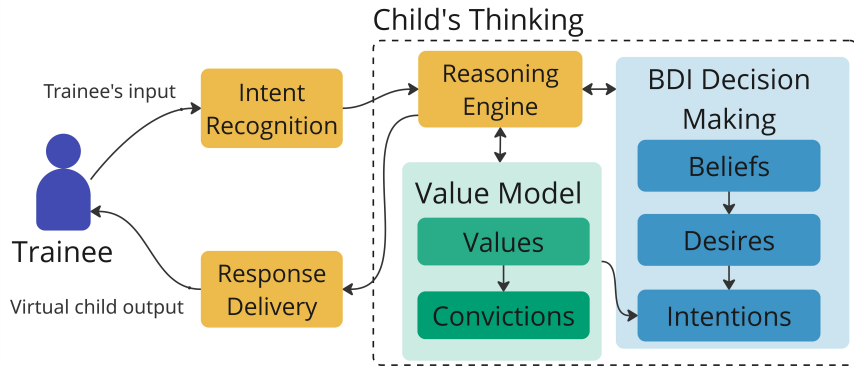


Figure 4.1: The BDI-values integrated model, based on the ARTES architecture [8]

4

4.3.1 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 [119] in the virtual child's thinking, wherein the child experiences discomfort from the mismatch between their behaviour and core values, increasing motivation to adopt actions that better reflect what they care about. Table 4.3 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) [261]. To build the system, we used an adolescent-oriented MI guidelines by Naar and Suarez [276], 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 behaviour 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 towards it). In the trainee's intent recognition, we modelled the four processes separately, with each MI process containing 4 to 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 behaviour, which supports progress towards change.

4.3.2 OVERALL BDI-VALUE MODEL

Before introducing values, we briefly clarify the underlying BDI mechanism used in Lilobot [145]. 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.

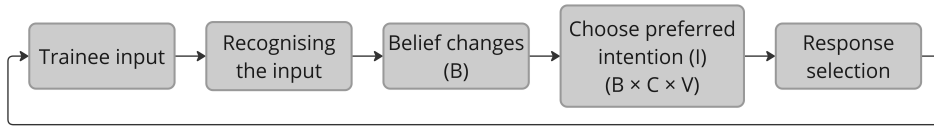


Figure 4.2: Single decision cycle in the virtual child’s BDI-based reasoning process. Each trainee’s input triggers intent recognition and then belief updates, after which the child re-evaluates its intention (action) before generating a response. This cycle is repeated for each trainee input.

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 4.1 shows the overall system flow with integrated values, adapted from the ARTES architecture for agent-based social skills training systems [8]. When a trainee inputs a message, the system recognises it as a 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,” where both have the active desire of “Lilo wants to be strong.” Lilo’s response is then selected from predefined 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 4.1 are explained in Table 4.1. To give an intuition about the value-integrated BDI process, Figure 4.2 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 4.3.4.

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. [257] 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 [336]. 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

Table 4.1: 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 modelled 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. [145] and Al Owayyed et al. [6]. 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.

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 prioritise 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 D.1.

4.3.4 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 operationalise 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 maximises 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 4.2:

$$\begin{aligned} \text{Hitting (H)} &= (B_1 \times C_1 \times V_{\text{Power}}) + (B_2 \times C_2 \times V_{\text{Security}}) \\ \text{Hitting (H)} &= (0.9 \times 0.8 \times 0.8) + (0.5 \times 0.5 \times 0.7) \\ &= 0.576 + 0.175 = \mathbf{0.751} \end{aligned}$$

$$\begin{aligned} \text{Walking Away (W)} &= (B_3 \times C_3 \times V_{\text{Security}}) + (B_4 \times C_4 \times V_{\text{Power}}) \\ \text{Walking Away (W)} &= (0.4 \times 0.5 \times 0.8) + (0.5 \times 0.6 \times 0.7) \\ &= 0.16 + 0.21 = \mathbf{0.37} \end{aligned}$$

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

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

Table 4.2: 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 4.3). Lilo also holds additional beliefs, desires, and intentions (Appendix D.1) 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)	

4.3.5 ILLUSTRATIVE CONVERSATIONAL EXAMPLE

Table 4.3: 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. Appendix D.5 includes conversation scripts of participants interacting with the virtual children during the experiment.

#	Dialogue	Belief Change	Intention Values	Preferred Intention
1	TRAINEE: What do you want to achieve in this conversation?	Initial: B1= 0.9	Hit =0.751 Walk=0.37	Hitting back
2	LILO: <i>I don’t want them to bully me.</i>	B2= 0.5		
3	LILO: <i>I want to show my classmates who is stronger.</i>	B3= 0.4 B4= 0.5		
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>			

¹Where *H* = Hitting and *W* = Walking Away.

#	Dialogue	Belief Change	Intentions	Preferred Intention
8	TRAINEE: So you might get in trouble if you fight.	B2↓↓ → 0.1	Hit =0.611 Walk=0.37	Hitting back
9	LILO: Yesss.			
10	TRAINEE: What other ways could you show you're strong?	B1↓↓ → 0.6 B2↓ → 0.0	Hit =0.512 Walk=0.494	Hitting back
11	LILO: I can walk away, maybe they will stop.	B3↑ → 0.5 B4↑ → 0.7		
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: Then they might say Lilo doesn't care about us.			
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: When they tease me, I will act like this doesn't bother me.			
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 will feel better and stronger.			

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.

Table 4.3 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 4.2) 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 recognises 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 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 favour of this new intention. In other words, this dialogue sequence illustrates how the child begins to recognise 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).

4.3.6 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 behavioural shifts throughout the interaction, without explicitly mentioning the values. We modelled 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.4).

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 behaviour 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..."

4.4 EVALUATION

We conducted a between-subjects experiment with three conditions to test the hypotheses presented in Section 4.1. We randomly assigned participants to one condition: (1) a virtual child with no explicitly modelled values (base BDI; taken from [6, 145]), (2) a virtual child with modelled values (BDI and values), or (3) a virtual child with modelled 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 motivational interviewing and the five-phase model to counsel the virtual children. Each experiment took approximately 75 minutes 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 Framework². All data, R analysis code (version 4.3.1), and a Docker container for reproducibility are publicly available via 4TU research data repository³. The system code is also publicly available on Github⁴.

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 [88].

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 4.4 shows a summary of the participant demographics, and Figure 4.3 presents the average self-reported value profile per condition. Based on these data, we observed no clear differences between the conditions.

4.4.2 MEASURES

PRIMARY MEASURES.

We used four primary measures to answer the hypotheses. All questionnaires are included in Appendix D.3.

Recognised child values. After each interaction, participants rated the virtual child's values using the Human Values Scale [335]. 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 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 difference 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).

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>

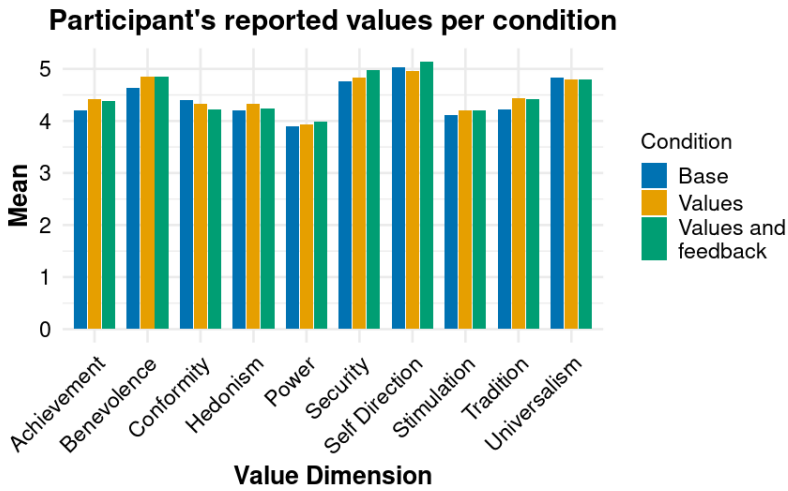


Figure 4.3: Participant's average ratings on their own values per condition.

Situational awareness. We measured participants' understanding of the child's values in a situation using the Situation Awareness Global Assessment Technique (SAGAT) [112, 328]. 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 [172] 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)[124, 125], but it too emphasises 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

⁵In 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.

Table 4.4: Summary of participant demographics by condition (Base BDI, BDI with values, BDI with values and feedback).

Demographic
Gender
<i>Base:</i> Man: 32 (50%), Woman: 31 (48.4%), Non-binary: 1 (1.6%)
<i>Values:</i> Man: 32 (50%), Woman: 30 (46.9%), Non-binary: 1 (1.6%), Prefer not to say: 1 (1.6%)
<i>Values-Feedback:</i> Man: 38 (58.5%), Woman: 26 (40%), Prefer not to say: 1 (1.5%)
Age Average
<i>Base:</i> 37.42 (13.9)
<i>Values:</i> 35.69 (12.7)
<i>Values-Feedback:</i> 34.97 (12.5)
Education Level
<i>Base:</i> Bachelor's: 29 (45.3%), Master's: 29 (45.3%), PhD: 4 (6.2%), Community college: 2 (3.1%)
<i>Values:</i> Bachelor's: 36 (56.2%), Master's: 22 (34.4%), PhD: 3 (4.7%), Community college: 3 (4.7%)
<i>Values-Feedback:</i> Bachelor's: 33 (50.8%), Master's: 21 (32.3%), PhD: 7 (10.8%), Community college: 4 (6.2%)
Country
<i>Base:</i> South Africa: 23 (35.9%), United States: 19 (29.7%), United Kingdom: 11 (17.2%), Others: 11 (17.2%)
<i>Values:</i> South Africa: 24 (37.5%), United States: 15 (23.4%), United Kingdom: 8 (12.5%), Others: 17 (26.6%)
<i>Values-Feedback:</i> South Africa: 28 (43.1%), United States: 13 (20%), United Kingdom: 11 (16.9%), Others: 13 (19.9%)
Has Children
<i>Base:</i> Yes: 49 (76.6%), No: 14 (21.9%), Prefer not to say: 1 (1.6%)
<i>Values:</i> Yes: 45 (70.3%), No: 19 (29.7%)
<i>Values-Feedback:</i> Yes: 47 (72.3%), No: 18 (27.7%)

Note. All demographic details are available in Appendix D.2.

inspired by the performance expectancy construct from the UTAUT model [383], and a prior questionnaire using this construct [123]. 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.”

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 Artificial Social Agent Questionnaire (ASAQ) [124, 125]. 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.”

4.4.3 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, summarising the five-phase model and motivational interviewing. For the BDI with values and value-with-feedback conditions, we used the virtual children model explained in Section 4.3. For the base BDI condition, we used two virtual children that did not explicitly model values, which were taken from the original system [6, 145]. 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.

4.4.4 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 motivational interviewing, 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 randomised. Each conversation had a 13-minute 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 recognised values. Next, participants were redirected to the system for a second 13-minute 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.

4.4.5 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 [217]. We also assessed internal consistency for the opportunity to align and perceived usefulness measures; the former showed lower reliability (Cronbach's $\alpha = .66$), whereas the latter showed high reliability ($\alpha = .92$). For all measures except the child's recognised 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 [76] and Andraszewicz et al. [16]. 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 [266] for Bayesian *t*-tests, the `brms` package [67] for multilevel modelling, and the `bayesestR` package [242] to calculate HDIs.

4.4.6 RESULTS

Here, we report the analysis results of the measures. Table 4.5 presents the means and standard deviations for all measures across conditions, while Table 4.6 summarises 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.

PRIMARY MEASURES

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

Recognised Child Values (H1) Participants recognised values most in the values-and-feedback condition. Figure 4.4 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 4.5, which are higher in the two value-based conditions. Statistical analysis indicates some support for increased value recognition in the values-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 ($M = 72.57$,

Table 4.5: Descriptive statistics (means, standard deviations, and 95% Highest Density Intervals (HDIs)) for each condition and hypothesis.

Hypothesis	Condition	Mean (SD)	95% HDI
Recognised 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]

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 4.6). 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

Table 4.6: Summary of the Bayesian test results, including posterior probabilities, 95% highest density intervals (HDIs), and posterior means and standard deviations for the group difference. The interpretations are based on the guidelines by Chechile [76] and Andraszewicz et al. [16].

Hypothesis	Comparison	Posterior Probability	Interpretation	Mean [95% HDI] (SD)
H1:	V > B	0.69	Not worth betting on	0.06 [-0.14, 0.25] (0.12)
Recognised	VF > B	0.9	A promising but risky bet	0.16 [-0.04, 0.36] (0.12)
Child Values	VF > V	0.8	Only a casual bet	0.10 [-0.10, 0.30] (0.12)
H2:	V > B	>0.99995	Virtually certain	1.29 [1.04, 1.55] (0.13)
Situational	VF > B	>0.99995	Virtually certain	1.42 [1.16, 1.67] (0.13)
Awareness	VF > V	0.78	Only a casual bet	0.12 [-0.2, 0.44] (0.16)
H3:	V > B	0.9996	Nearing certainty	0.35 [0.16, 0.52] (0.09)
Opportunity	VF > B	0.9999	Nearing certainty	0.36 [0.17, 0.55] (0.10)
to Align	VF > V	0.54	Not worth betting on	0.01 [-0.19, 0.20] (0.10)
H4:	V > B	0.76	Only a casual bet	0.11 [-0.20, 0.40] (0.15)
Perceived	VF > B	0.87	Only a casual bet	0.18 [-0.12, 0.5] (0.16)
Usefulness	VF > V	0.69	Not worth betting on	0.08 [-0.23, 0.38] (0.15)
Secondary:	V > B	0.97	Good bet-too good to disregard	0.53 [-0.04, 1.08] (0.29)
Believability	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:	V > B	0.95	Good bet-too good to disregard	0.33 [-0.08, 0.72] (0.20)
Overall	VF > B	0.89	Only a casual bet	0.25 [-0.14, 0.66] (0.20)
Experience	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. *Note.* 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 values-and-feedback condition to the base condition yielded a posterior probability of 0.9999 (95% HDI [0.17, 0.55]), nearing certainty in favour 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

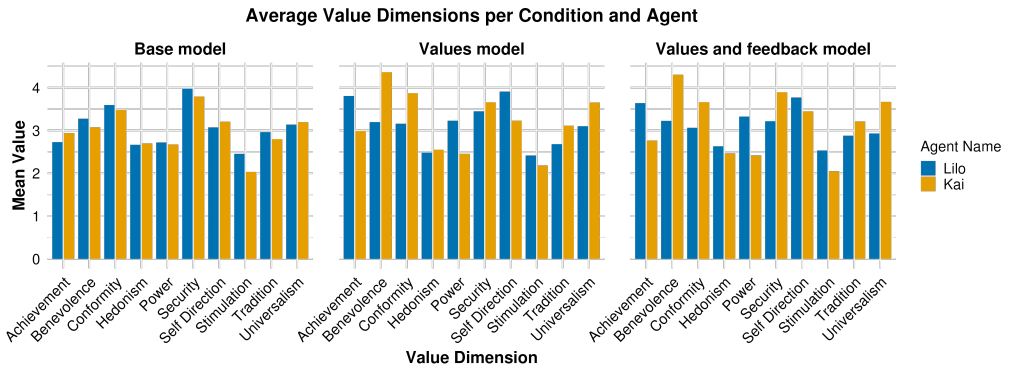


Figure 4.4: How participants rated the values in each agent (Lilo and Kai), across the three conditions.

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: $M = 6.06$).

SECONDARY MEASURES

Like the primary constructs, the secondary measures showed general trends favouring modelled 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-modelled conditions (posterior probability = 0.3) provides no support. Thus, adding values to the virtual child simulation appears to improve believability.

Overall Experience with the Agent Figure 4.5 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 utilised the same questionnaire [125]. Figure 4.6 shows that the two values conditions exceed 95% of the set, while the base condition exceeds the 75% mark.

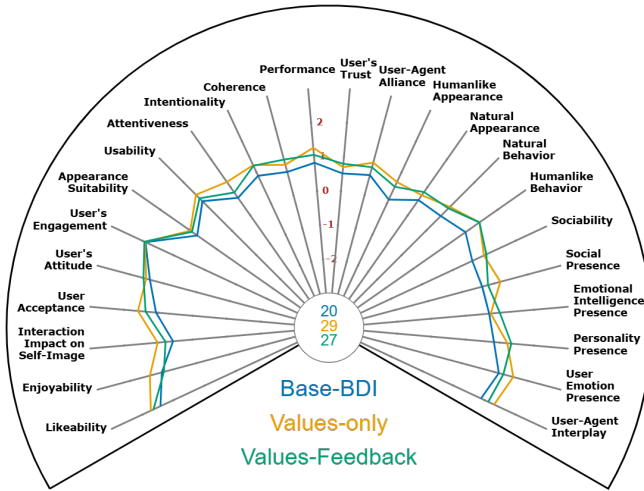


Figure 4.5: Comparison between the three conditions across all ASAQ constructs. The chart was generated using the official ASAQ visualisation tool.

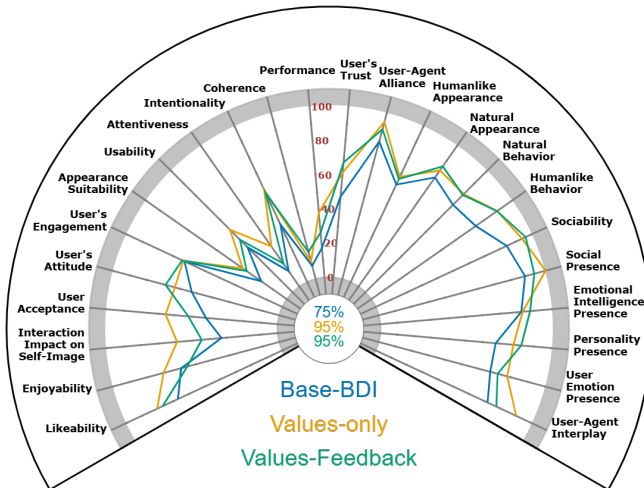


Figure 4.6: Percentile chart comparing the three agents to the ASAQ Representative Set 2024 [125]. The grey area indicates the performance range of the 29 agents in the benchmark dataset. The chart was generated using the official ASAQ visualisation tool.

4.5 DISCUSSION AND CONCLUSIONS

4.5.1 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 counsellors at helplines.

The results indicate that the training system can help trainees recognise (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 [384].

Generally, participants' identified values matched with the intended values we modelled 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 [37, 150].

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 [229]. In e-learning systems, users' motivation can directly shape how they engage with and accept a system [114]. According to the Technology Acceptance Model (TAM), three key factors affect people's motivation to use such systems: perceived usefulness, perceived ease of use, and attitude toward using them [95, 142]. 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. [45], 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 favour 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 favoured the opposite. Comparing overall experiences using the short-ASAQ percentile ratings (Figure 4.6), all three agents scored highly on constructs related to sociality, such as natural behaviour, 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.

4.5.2 LIMITATIONS AND FUTURE WORK

One limitation of this study is that we evaluated the agent's ability to make children's values recognisable and to provide opportunities for alignment, rather than assessing whether the training improved counselling skills, such as the ability to apply MI. Evaluating the impact on counselling skills would require more than a single session, as traditional MI training can take several days to complete [276]. For example, in a non-randomised controlled trial, training counsellors in MI at a suicide helpline produced similar outcomes to training on the five-phase model [182]. 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 modelled 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 D.5.

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 behaviour, their interpretation is inherently subjective [139, 280], which may contribute to variability in situational awareness assessments. Different raters may emphasise 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 re-examine 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 .70 threshold for acceptable reliability [41, 285]. To examine the robustness of the findings, we removed two items (items 4 and 5; Appendix D.3), which resulted in acceptable reliability ($\alpha > .70$). 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-values model [7]. Additionally, the experiment was limited to two scenarios focusing on four values from Schwartz's model, which may limit the generalisability 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 [120, 365], communities [149], and trainees themselves [298]. The latter is particularly relevant in cases of value conflict during counselling, where discrepancies between a counsellor's and a client's values may arise. Several frameworks and strategies exist for managing such conflicts (e.g., [116, 203]), 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.

4

4.5.3 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 behaviours associated with those values, which could risk reinforcing stereotypes in training systems [8]. Therefore, training should incorporate a variety of perspectives and virtual child behaviours to avoid associating specific values with action. Trainees should also be mindful of the potential for unethical use of MI to influence behaviour, and must adhere to ethical guidelines [261, 276].

Our training simulation prioritises internal value trade-offs 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 counsellors to challenging scenarios. At the same time, deciding which values are 'good' or 'bad' involves many intersecting factors that make moral judgment complicated [141]. 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.

4.5.4 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 counsellors for value-sensitive conversations.

5

INTEGRATING LLMs WITH A BDI-BASED CONVERSATIONAL AGENT FOR CHILD HELPLINE TRAINING

5

Child helpline training often relies on human-led roleplay, which is both time- and resource-consuming. To address this, rule-based interactive agent simulations have been proposed to provide a structured training experience for new counsellors. However, these agents might suffer from limited language understanding and response variety. To overcome these limitations, we present a hybrid interactive agent that integrates Large Language Models (LLMs) into a rule-based Belief-Desire-Intention (BDI) framework, simulating more realistic virtual child chat conversations. This hybrid solution incorporates LLMs into three components: intent recognition, response generation, and a bypass mechanism. We evaluated the system through three studies: (1) a script-based assessment comparing LLM-generated responses to human-crafted responses, (2) a within-subject experiment ($N = 37$) comparing the LLM-integrated agent with a rule-based version, and (3) interviews with three helpline experts who tried the system and were then interviewed about how LLMs are integrated in training contexts. The first study provided evidence that the three LLM components were non-inferior to human-crafted responses. In the second study, we found credible support for two hypotheses: participants perceived the LLM-integrated agent as more believable and reported more positive attitudes toward it than the rule-based agent. Additionally, although weaker, there was some support for increased engagement (posterior probability = 0.845, 95% HDI $[-0.149, 0.465]$). The third

This chapter is based on a previously published paper, with the addition of a third study involving experts. Accordingly, the abstract, introduction and overall discussion were revised, and a new section for Study 3 was added. The original paper is:

📖 **Al Owayyed, M., Denga, A., and Brinkman, W. P.** (2025, September). Controlled Yet Natural: A Hybrid BDI-LLM Conversational Agent for Child Helpline Training. *In Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents* (pp. 1–10).

study identified key training-related considerations, including the importance of modelling teachable moments and the child's internal states, as well as incorporating feedback. Our findings demonstrate the potential of integrating LLMs into rule-based systems, offering a promising direction for more flexible but controlled training systems.

5.1 INTRODUCTION

Child helplines offer a safe space for children in need to seek help. This requires extensive counselling training efforts. Children can contact helplines such as De Kindertelefoon through online chat interfaces or by phone. As of 2023, the Dutch Children's Helpline, De Kindertelefoon, has over 700 volunteer counsellors, with more than 340 new volunteers trained that year—an increase of 20% compared to 2022 [97]. To train new counsellors, a common practice is role-playing, where one counsellor acts as a child contacting the helpline while the trainee engages in conversation. However, this method can be time-consuming and resource-intensive. A potential alternative is using interactive social agents to simulate child-helpline interactions, which were helpful in similar training scenarios, such as virtual patients in medical education [146, 401].

A crucial aspect of a simulation system for learning is being as faithful to the actual situation [329]. In this case, simulating realistic, open-ended chat conversations of a virtual child. However, training counsellors using virtual agents can be sensitive, especially when using open-ended unrestricted interaction [38]. While closed-answer options provide a safer environment, they do not fully replicate real-world interactions where counsellors formulate responses independently to develop their chat conversational skills.

One way to enable open-ended conversations is through rule-based intent recognition, where the system interprets user input and selects appropriate responses. An example of such a system is Lilobot, a BDI-based (Belief-Desire-Intention) training system designed for child helplines [6, 145]. The system focuses on bullying scenarios, one of the main topics children discuss with De Kindertelefoon [200]. Lilobot simulates a virtual child who has been bullied at school and is seeking support from a helpline. The system follows a five-phase model of counselling conversations [344], which trainees should apply correctly for the interaction to progress. The phases are building rapport, clarifying the child's story, setting the conversation goal, working towards that goal and wrapping up the conversation. If a counsellor fails to follow these phases, Lilobot may leave the conversation.

However, traditional intent recognition has limitations, as it can only respond within the constraints of its predefined rules and training data [311]. This was also reported for Lilobot. Although positively received for its training potential, participants highlighted certain limitations. Specifically, the system struggled to understand complex sentences outside its knowledge base, indicating a need for improved comprehension. Participants also expressed a need for more response variations, as variety and realism are important for effective learning [23]. These challenges align with the broader limitations of rule-based conversational agents [311].

Recent progress in large language models (LLMs) offers a promising approach for overcoming these limitations. LLMs can generate human-like text with contextual relevance and variability, making conversations appear more natural and engaging compared to natural language understanding and generation [128, 147, 192]. However, using pure LLMs in training scenarios presents a challenge; while they enhance realism, they should operate

within a controlled learning environment to ensure specific educational objectives are met. Without appropriate constraints, LLMs may generate unintended or extraneous content, which could interfere with the learning process, adding extraneous load to the trainee.

To balance these considerations, we propose integrating LLMs with Lilobot's existing BDI framework. This paper presents an integration design that preserves the structured learning environment required for counsellor training and evaluates its effects. Our evaluation includes three studies: (1) a non-inferiority analysis, conducted through human coders, comparing LLM-generated and BDI system responses in terms of both understanding and response quality; (2) an experimental study where participants used both LLM-integrated and BDI-only systems in a within-subject design; and (3) interviews with helpline experts who reviewed the system, aimed at identifying training-relevant considerations regarding the integration of LLMs. Based on the second study, we formulate and test the following hypotheses:

H1: Individuals perceive the integrated system to be more believable than the rule-based system.

H2: Individuals perceive the integrated system to be more engaging than the rule-based system.

H3: Individuals have a more positive attitude towards the LLM-integrated system than the rule-based one.

5.2 RELATED WORK

5.2.1 SOCIAL AGENTS FOR HELPLINE TRAINING

Researchers explored interactive social agents as tools for training helpline counsellors. Demasi et al. [99] proposed a chatbot simulation to train counsellors in suicide prevention hotlines. This simulation uses annotated transcripts and a multi-task framework to generate responses that mimic various crisis conversation scenarios. Evaluations of the proposed model showed improvements in response diversity and specificity compared to simpler models. For child helpline training specifically, one example is a serious game designed to simulate high-risk interactions in a safe environment (e.g., trafficking and sexual exploitation) [382]. This game uses choice-based interactions to train counsellors in handling sensitive situations while developing essential skills. However, it does not support open-ended interactions, which helps avoid language processing and generating limitations.

5.2.2 LLMs FOR COUNSELLING

LLMs have been used to support counselling in various ways. To simulate a counsellor providing help, Steenstra et al. [354] explored LLMs' potential in delivering motivational interviewing for alcohol counselling. They found that LLM-powered virtual agent responses were perceived similarly to human-generated ones, addressing the limitations of rule-based approaches in understanding nuanced therapy conversations. Recently, Heinz et al. [162] reported a reduction of symptoms in depression, anxiety, feeding and eating disorders after people interacted for four weeks with their therapy chatbot Therabot. This text-based application used a generative LLM fine-tuned on expert-curated mental health dialogues. For training counsellors, Wang et al. [387] introduced an LLM-driven virtual patient to

train trainees in formulating cognitive models as part of cognitive behavioural therapy. The authors defined several cognitive models as scenarios and instructed the LLM to simulate patient behaviour using them. This also allows trainees to compare the identified cognitive model with the actual one. When evaluating the system, the simulated interactions resulted in greater perceived knowledge and confidence compared to traditional methods. Other examples of LLM-integrated virtual patients include applications in history taking [169], and medical diagnosis [331, 381]. LLMs have also been used to support counsellors in their work or training by generating reflections [339], providing feedback on trainees' responses [63, 75], and offering suggestions to counsellors [215].

5.2.3 INTEGRATIONS OF LLMs AND BDI

The integration of LLMs with rule-based systems has been explored to enhance simulations' reliability, realism, and explainability. Pico et al. [304] examined integrating LLMs with a BDI model to improve emotion recognition in intelligent agents. In their approach, an LLM is prompted with a dialogue and tasked to pick up an emotion, which is then converted into beliefs used by the BDI model. When evaluated, the LLM demonstrated promising capabilities in emotion recognition. Frering et al. [129] integrated LLMs and BDI in a human-robot interaction setting to improve explainability. The authors used an LLM to interpret user input and translate it into commands the BDI agent could process. Additionally, the LLM generated responses explaining the robot's behaviour based on the BDI agent's state.

5.3 INTEGRATING LLMs WITH LILOBOT

We integrated LLMs into Lilobot's rule-based framework to enhance understanding and generation capabilities while maintaining control over scenario structure. We examined integration possibilities within the existing system, explored LLM models and prompt strategies, and then selected a practical implementation approach.

5.3.1 INTEGRATION POSSIBILITIES

Lilobot follows a BDI architecture [145], which defines internal states for the virtual child, making its decision-making controllable and interpretable. The system maps trainee input to predefined intents—for example, when a trainee writes to the virtual child, “How does that make you feel?”, the input is recognised as the intent “request_unknown_feeling.” These intents are mapped to child beliefs within the BDI framework, which are tied to the child's desires and intentions—for instance, the intent will increase the belief that “they feel the trainee can be trusted.” Based on the current BDI state and the recognised trainee intent, the system picks a response from its knowledge base. To introduce variation, each response has four alternatives, one of which is selected at random; in this case, the child might reply, “It makes me sad... I really don't know what to do.” If the trainee's intent is not recognised, the system retrieves a default response based on Lilobot's currently active desire (i.e., goal). For example, when the desire “they want the trainee to call their school” is active, a default response could be “I want you to call my school.” While this structure ensures consistency, it also limits the dialogue's flexibility and adaptability to diverse responses.

To identify where we can integrate LLMs, we followed the general architecture of agent-based training for social skills (ARTES) [8] and Lilobot’s limitations. ARTES outlines three possible integration points in the simulation: Natural Language Understanding (NLU), Natural Language Generation (NLG), and Lilobot’s thinking. NLU and NLG have been shown to perform well with LLMs [74], allowing for a better understanding of trainee input as well as increased variability and naturalness in responses. As for the thinking component, integrating an LLM would enable Lilobot to bypass unexpected cases not covered in the list of trainees’ intents. Therefore, we refer to it as the Bypass component in this paper.

5.3.2 COMPONENTS

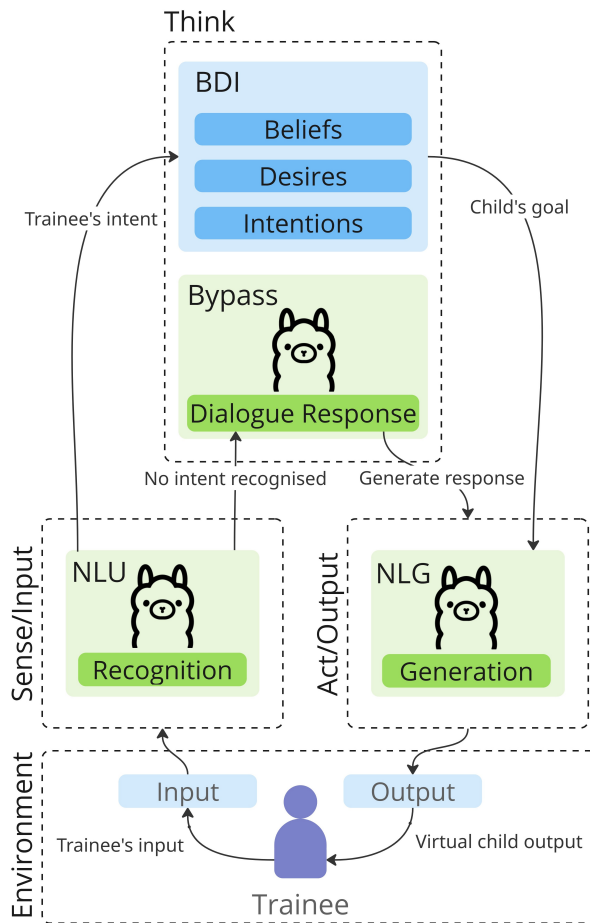


Figure 5.1: Architecture of the LLM-integrated BDI-based conversational agent based on the ARTES architecture [8]. The green components indicate where the LLM is integrated.

Figure 5.1 shows the integrated system’s architecture. The overall flow begins with the

trainee sending an input, which the NLU component matches to a trainee intent using over 2,000 examples from a dataset used by the rule-based system. If a match is found, the intent is sent to the BDI system to update the cognitive state of the virtual child, resulting in the identification of a response with four associated example responses. These are passed on to the NLG component along with the original input message from the trainee. The NLG component, in turn, generates a contextually appropriate response for the trainee. If the NLU finds no matching intent, the Bypass component generates a relevant response based on the trainee’s input, which is then sent back to the trainee.

5.3.3 PROMPT CREATION

We explored several prompts for the three components: NLU, NLG, and Bypass. We followed prompt engineering techniques [21, 370]. This led to specific prompt designs, one for each component.

NLU: To classify trainee input into a BDI-recognisable intent, we designed a prompt in which the LLM’s task is to select the best-matching intent for the input or return “*unknown*” if no match is found. When a trainee input is received, it is embedded and compared—using L2 distance in embedding space—to over 2,000 annotated examples (e.g., “*why are you being bullied?*” mapped to the intent *bullying_why*), which are stored in a vector database. The closest examples to the trainee’s input are then included in the prompt, along with their corresponding intents and the trainee’s input itself, and sent to the LLM. This method reduced the prompt size while preserving contextual relevance. In our initial attempts, we embedded all examples in the prompt, which resulted in slower response times.

NLG: When a trainee’s input is matched to an intent, it is sent to the BDI in the Think component, along with the child’s state, and then passed to the NLG sub-component to generate a response. The NLG prompt includes the trainee’s input, a defined child persona (e.g., “*You must play the character of Lilo, a 9 year old child being bullied at school.*”), the child’s current goal as represented in the BDI model, and four example child responses linked to the identified trainee intent. The prompt instructs the LLM model to generate a reply similar in tone and structure to these examples. The resulting response is then sent to the trainee as the virtual child’s reply.

Bypass: When no matching trainee intent is found, the Think component triggers the Bypass sub-component. The prompt structure is similar to that of the NLG prompt, with the key difference being the absence of example responses. In this case, the LLM is instructed to generate a contextually appropriate response from a child’s perspective without relying on predefined examples. The generated response is then sent to the trainee.

5.4 STUDY 1: SCRIPT EVALUATION

To evaluate the effectiveness of the prompts, we assessed whether the outcomes generated by the LLM were non-inferior to those produced by human-crafted responses. These human-crafted responses, drawn from Lilobot’s rule-based model, had been reviewed by domain experts [145]. Inspired by Steenstra et al. [354], we employed a within-subjects design in which we asked four human coders to rate or label the outcomes of both the LLM prompts and the human-crafted responses from the rule-based system. The coders included an undergraduate medical student, a computer science graduate, and two master’s

students in computer science. For NLU, we focused on categorisation accuracy. We created 12 trainee input examples, where the LLM prompt, the rule-based system, and the coders independently matched each utterance to a trainee intent, or labelled it as "unknown" if it did not match any trainee intent. For NLG and bypass, we evaluated whether the language used in the LLM-generated responses was appropriate to the context, the virtual child's state, and the trainee's input. We generated eight NLG responses per system: eight human-crafted responses from the rule-based system and eight LLM-generated counterparts. Additionally, we generated eight LLM bypass responses that could not be matched to intents in the rule-based system. Participants were then presented with all 24 responses in randomised order, along with the conversation history for context, and were asked to rate them. The source of each response was concealed from the participants. All data, prompts, pipelines for the prompts, and analysis code are available online ¹.

5.4.1 MEASURES

NLU: Participants were given a list of 38 trainee intents with brief explanations and 12 trainee inputs. Their task was to match each input to one or more intents. In the ground truth, 10 inputs were each mapped to a different intent, and the remaining two were mapped to two intents each, resulting in 14 intents to identify.

NLG and Bypass: Participants were asked to rate each response on a Likert scale (1 = Strongly Disagree to 7 = Strongly Agree) based on five statements adapted from Steenstra et al. [354] (Table 5.1).

Table 5.1: Rating statements for the Bypass and NLG evaluation tasks, adapted from [354].

Statements
C1: This response is in coherent English
C2: This response is coherent in this bullying context
C3: This response directly addresses and replies to the counsellor's previous message
C4: This response makes sense
C5: This response makes sense in this context of bullying

5.4.2 DATA ANALYSIS

We conducted a Bayesian analysis to perform a non-inferiority test. Specifically, we fitted a Bayesian multilevel regression model, using non-informative priors. For NLU, we used the model to estimate whether the intent categorisation result for each example matched the ground truth using the rule-based model or the LLM. For the NLG and Bypass components, the model accounted for variability across individual coders when rating LLM- and human-crafted utterances. In both cases, we defined a non-inferiority threshold just below the null value. We then assessed the proportion of the posterior distribution of the parameter estimate that fell above this threshold, which would indicate practical non-inferiority [214]. The threshold was set based on half of Cohen's [271] threshold for a small effect size ($-0.1 \times \text{sd}(\text{score})$) to define negligible or better effects [214]. Additionally,

¹<https://data.4tu.nl/datasets/5c5b69b7-f727-42d0-983f-07ab910b8460>

inter-rater reliability was assessed using the Intraclass Correlation Coefficient (ICC), Fleiss' and Cohen's Kappa to estimate agreement between coders, the LLM, the rule-based system, and ground truth.

5.4.3 RESULTS AND DISCUSSION

In the NLU analysis, the LLM correctly recognised all 14 trainee intents for the 12 examples, while the rule-based system recognised 12 intents, and the median of the four human coders recognised 13. To check for overall reliability, Fleiss' Kappa among human coders alone was 0.81, and between the coders and the LLM was 0.82, which Landis and Koch [217] classify as "almost perfect" agreement. Agreement between coders and the rule-based model was slightly lower but still substantial ($\kappa = 0.78$), according to the same classification scale. Cohen's Kappa indicated perfect agreement between the LLM and ground truth ($\kappa = 1.00$), and almost perfect agreement between the median coder and ground truth ($\kappa = 0.92$), as well as between the rule-based model and ground truth ($\kappa = 0.85$) [217]. The non-inferiority results indicated some support for the non-inferiority of the LLM, with the probability of 0.87 being non-inferior.

For NLG and Bypass, coders' ratings agreement was generally low ($ICC < 0.5$). Median ratings clustered near the top of the scale (coder 1 = 6.5, coder 2 = 7, coder 3 = 7, coder 4 = 5.4), potentially indicating ceiling effects and limited variance. LLM texts received a mean rating of 6.30 ($SD = 0.89$), while human-crafted texts had a mean of 6.28 ($SD = 0.78$). The results showed a posterior probability of 0.70, indicating some support for the non-inferiority of LLM-generated content relative to human-crafted text.

5.5 STUDY 2: EXPERIMENT

After evaluating the LLM-generated script quality, we conducted a within-subjects experiment to test our hypotheses. Participants role-played as counsellors tasked with applying the five-phase model with the virtual child. There are two conditions: (1) a rule-based agent and (2) an LLM-integrated agent. Each participant experienced both systems in a randomised order to control for order effects, with the entire experiment lasting approximately 50 minutes. This experiment was approved by the university human research ethics committee (HREC reference number: 4768) and preregistered with the Open Science Framework². The data, analysis code in R (version 4.3.1), and a Docker container for reproducibility are publicly available via the 4TU research data repository³.

5.5.1 PARTICIPANTS

We recruited participants through the online platform Prolific. To match the target group for this training system, e.g., new helpline volunteers at organisations like De Kindertelefoon, we determined eligibility as adults with at least a Bachelor's degree. We also required fluency in English, an approval rating of $\geq 95\%$, and have completed at least five studies in the past. A total of 55 participants were invited; 18 were excluded in total due to incomplete participation ($n = 13$), failing comprehension or attention checks ($n = 2$), not providing consent ($n = 1$), or not meeting screening criteria ($n = 2$). This resulted in a final sample of

²<https://osf.io/eqxwz>

³<https://data.4tu.nl/datasets/5c5b69b7-f727-42d0-983f-07ab910b8460>

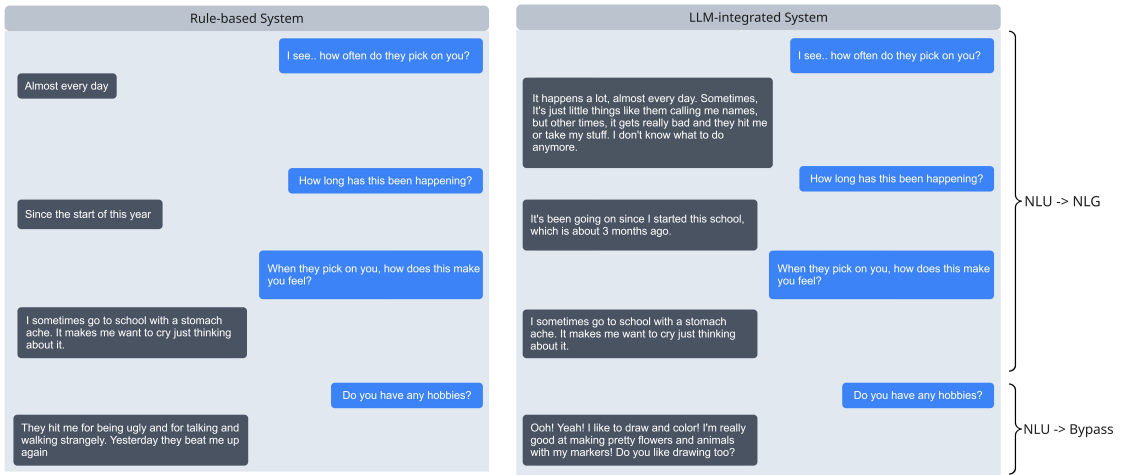


Figure 5.2: Example conversation for each condition. The rule-based system (left) and the LLM-integrated system (right) respond differently to the same counsellor input. The last input does not exist in the trainee intents list, triggering the Bypass component in the LLM-integrated system. Meanwhile, the rule-based system responds with a message related to the child’s active desire.

37 participants. The study was conducted between January and February 2025. Participants received compensation per Prolific’s minimum payment policies.

Of the 37 participants, 24 held a Bachelor’s degree or equivalent (65%), 10 held a Master’s degree (27%), and 3 held a PhD (8%). Fourteen participants identified as male (38%) and 23 as female (62%). Age distribution was as follows: 18–24 (n = 10), 25–34 (n = 12), 35–44 (n = 9), 45–54 (n = 5), and 55–64 (n = 1). Participants represented diverse nationalities, including the United Kingdom, Portugal, the United States, South Africa, Nigeria, and India.

5.5.2 MEASURES

We tested primary and secondary measures to evaluate the rule-based and integrated systems. The primary measures were used to test the three hypotheses, while the secondary measures provided additional insight into participants’ experiences.

PRIMARY MEASURES

We have three measures corresponding to the hypotheses: Believability, Engagement, and Attitude. All questionnaire items were drawn from the Artificial Social Agent Questionnaire (ASAQ) [124, 125], using a 7-point scale from -3 (Disagree) to +3 (Agree), with 0 as Neither agree nor disagree. We replaced references to “the agent” with “the virtual child” in all questionnaire items.

Believability: Measured using two constructs from the ASA questionnaire, Human-Like Behaviour (HLB) and Natural Behaviour (NB). They capture how believable the agent is to participants.

Engagement: Measured using the engagement construct, assessing participants’ involvement during the interaction.

Attitude: Measured using the attitude construct, reflecting participants’ feelings toward the virtual child.

SECONDARY MEASURES

To gain further insight, we included the following secondary measures:

Overall Experience: Assessed using the 24-item short version of the ASA questionnaire, capturing broader constructs evaluation, expressed in short ASAQ score.

Preference: At the end of the experiment, participants answered a preference question: “Which virtual child did you prefer interacting with?”, requiring a selection between “The first child” or “The second child”.

Experience: Qualitative feedback was collected through an open-ended question after each condition (“How did your interaction with the virtual child go?”).

5.5.3 SYSTEM

Both conditions (rule-based and LLM-integrated) were presented through an identical web interface hosted on TU Delft servers. The layout displayed the conversation on the right and a summary of the five-phase model on the left. Each participant interacted with two of the twelve bullying-related scenarios, with each virtual child having a different, randomly assigned name. Scenarios were randomised across conditions to avoid systematic scenario bias and were assigned such that no participant encountered the same scenario twice. As described above, the systems differed only in how they processed and responded to the participants’ input. The rule-based agent used Rasa for NLU and predetermined responses, while the LLM-integrated version used Llama 3.2, run via Ollama, for NLU, NLG, and bypass. To ensure a fair comparison, we modified the rule-based system to match the response time of the LLM-integrated system (a random time between 15-25 seconds). Figure 5.2 shows a conversational example from the rule-based system (left) and the LLM-integrated one (right). The code is publicly available online⁴.

5.5.4 PROCEDURE

Participants were redirected from Prolific to Qualtrics, where they completed an informed consent form. They then watched a short training video introducing the five-phase model and providing context on counselling practices. Afterwards, they interacted with the two systems (rule-based and LLM-integrated), with the order chosen at random. Participants were first directed to one of the systems and instructed to interact with it for 15 minutes. Within this time, the scenario restarted if the virtual child exited or if the participant ended the interaction (e.g., by saying “bye”). After 15 minutes, participants were redirected to Qualtrics to complete questionnaires on believability, engagement, attitude, overall experience, and an open-ended reflection question based on the virtual child they had just interacted with. They were then redirected to the second system, following the same structure: 15 minutes of interaction, followed by the same questionnaires based on the second virtual child. At the end, participants chose their preferred virtual child.

⁴<https://github.com/adarshdenga/llm-integration-childhelpline>

5.5.5 DATA ANALYSIS

We analysed data from the 37 participants who were included in the study. We reversed items where required for all questionnaire-based measures, according to the ASA questionnaire guidelines, and computed average scores per construct. To test the hypotheses, we conducted Bayesian⁵ paired-sample *t*-test using non-informative priors. We calculated the posterior probability that the difference between the averages (LLM-integrated – rule-based) is greater than zero and interpreted it based on the guidelines by Chechile [76] and Andraszewicz et al. [16]. We also checked whether zero is included in the credible intervals expressed by the Highest Density Interval (HDI). For the preference question, we used a Bayesian binomial test. We performed a thematic analysis [56] of the qualitative responses to identify recurring patterns and insights across the open-ended questions. To ensure reliability, two coders with a background in computer science independently coded the responses using a predefined coding scheme. The inter-coder reliability was fair (Cohen's $\kappa = 0.32$) [217]. A third coder then reviewed the disagreements to determine the final coding.

5.5.6 RESULTS

5

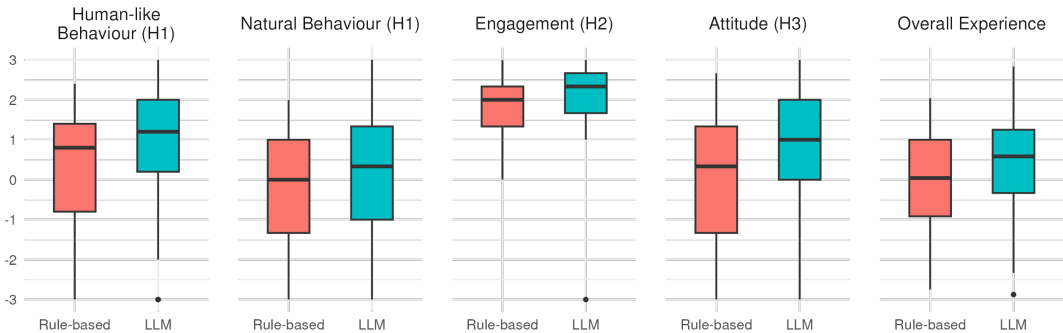


Figure 5.3: Comparison of participant ratings across five constructs between the rule-based and LLM-integrated agents.

PRIMARY MEASURES

Figure 5.3 shows an overall comparison across all measures. The overall trend seems to favour the LLM-integrated system.

⁵We reported the Bayesian analysis results in the main text to get more insight and to be consistent with Study 1. In the preregistration, we indicated using a frequentist test by specifying a cutoff value of $p < 0.05$. For transparency, the frequentist results for each measure are as follows:

- **Human-like Behaviour (H1):** $t(37) = 2.10, p = 0.04^*$
- **Natural Behaviour (H1):** $t(37) = 1.38, p = 0.18$
- **Engagement (H2):** $t(37) = 1.07, p = 0.29$
- **Attitude (H3):** $t(37) = 2.46, p = 0.02^*$
- **Overall Experience:** $t(37) = 2.57, p = 0.01^*$
- **LLM Preference:** $\hat{p} = 0.70, 95\% \text{ CI } [0.53, 0.84], p = .02^*$

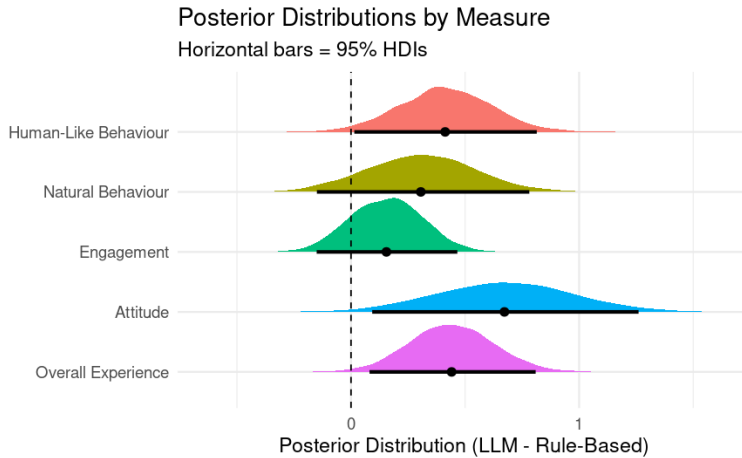


Figure 5.4: Posterior distributions of effect sizes comparing the LLM-integrated to the rule-based system across outcome measures.

5

Believability (H1) Human-like behaviour and natural behaviour were rated higher for the LLM-integrated system (HLB: $M = 0.79$, $SD = 1.62$; NB: $M = 0.14$, $SD = 1.55$) than the rule-based system (HLB: $M = 0.35$, $SD = 1.57$; NB: $M = -0.19$, $SD = 1.57$). The posterior probability for the human-like behaviour was 0.975, a good bet in favour of the LLM system, with a 95% HDI of [0.002, 0.816] which excludes zero. This further supports the difference as the HDI range is above zero (Figure 5.4). For the natural behaviour, the posterior probability was 0.905, which is considered a promising but risky bet for the LLM-integrated system. This is also shown in the 95% HDI [-0.161, 0.771], which overlaps with zero. Given these two results, there is credible support for H1.

Engagement (H2) Engagement ratings were slightly higher for the LLM-integrated system ($M = 2.01$, $SD = 1.05$) than the rule-based system ($M = 1.84$, $SD = 0.77$). The posterior probability of 0.845 constitutes only a casual bet, with 95% HDI [-0.149, 0.465] included zero (Figure 5.4); thus, there is some support for H2.

Attitude (H3) Participants reported a more favourable attitude toward the LLM system ($M = 0.86$, $SD = 1.52$) than the rule-based system ($M = 0.13$, $SD = 1.76$), with a posterior probability of 0.988—a good bet for the LLM system. The 95% HDI [0.083, 1.266] is above zero (Figure 5.4). This indicates credible support for H3.

SECONDARY MEASURES

Similar to the primary measure, the overall trend seems to favour the LLM-integrated system.

Overall Experience The overall experience rating is also higher for the LLM-integrated system ($M = 0.46$, $SD = 1.25$, Short ASAQ score = 11) than the rule-based system ($M = -0.021$, $SD = 1.28$, Short ASAQ score = -1). The posterior probability was 0.991—a strong bet and irresponsible to avoid, with the 95% HDI [0.075, 0.807] higher than zero (Figure 5.4). We also used the ASAQ Representative Set 2024, consisting of 29 artificial social agents,

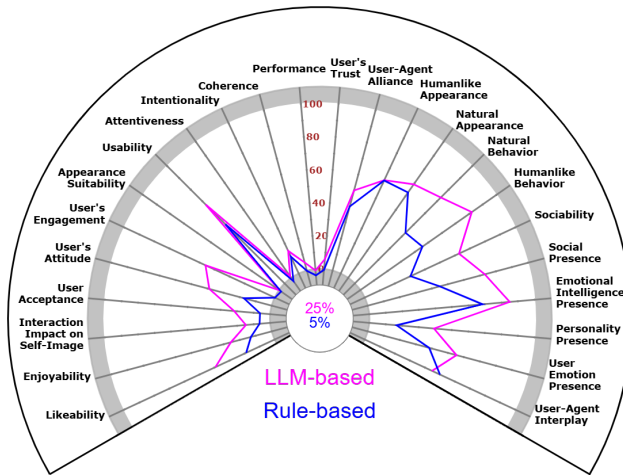


Figure 5.5: Percentile chart comparing the LLM-integrated and rule-based agents against ASAQ Representative Set 2024 [125]. The grey area represents the range of scores higher or lower than the 29 agents in the dataset. The visualisation was created using ASAQ’s visualisation tool [125].

5

for broader cross-study benchmarking, as suggested by the ASAQ’s authors [125]. This resulted in Figure 5.5, where the LLM-integrated agent outperformed 25% of the set, while the rule-based agent exceeded only 5%.

Preference The preference analysis indicates a preference for the LLM system, with 26 participants favouring it. The posterior probability that this preference is above chance is 0.99, which is a strong bet that the LLM-integrated agent is preferred.

Qualitative Feedback Table 5.2 presents nine identified themes and their frequency in each condition. The themes relate either to the child’s replies (the first five themes) or to the interaction experience (the last four). Overall, positive themes were more prevalent in the LLM condition, such as perceived conversational depth and human-like responses (e.g., “... I think it acted of of a human being. It was quite impressive”). In contrast, negative responses were more frequent in the rule-based condition, including unnatural and slow responses (e.g., “It was boring, non-responsive, had delayed responses, did not reveal any realistic feelings, etc.”). One theme observed in both conditions—abrupt endings—may be partly attributed to the controlled scenario design, in which the virtual child could exit the conversation if the trainee deviated from the five-phase model (e.g., failing to show empathy or directly suggesting an action).

5.5.7 DISCUSSION

In this study, we evaluated whether integrating LLMs into a rule-based system could enhance believability (H1), engagement (H2), and positive trainee attitudes (H3) toward a virtual child used for helpline counsellor training.

We found credible support for believability (H1), particularly human-like behaviour. Furthermore, we hypothesised greater engagement for the LLM-integrated system than the rule-based one (H2). The results provided some support for this. Although scores were

Table 5.2: Themes identified from participants’ responses and the frequency of their occurrence in each condition.

Theme	LLM-integrated	Rule-based
Abrupt Ending	3	7
Unnatural Responses	4	7
Depth of Conversation	4	0
Emotional Engagement	6	1
Human-Like Responses	4	2
Positive Experience	9	8
Scripted Responses	0	1
Slow Responses	6	10
Boring Task	1	1

5

slightly higher, the posterior distribution indicated only a casual probability in favour of the LLM system. This may be attributed to a ceiling effect, as both systems scored highly on engagement (Figure 5.3). Such high engagement may result from the inherently engaging nature of the task, that is, providing support to a child in need. Additionally, participant comments clearly indicated task-related engagement (e.g., for the rule-based system: “...we worked together to talk about their feelings and challenges at school. It was a productive conversation...”, and for the LLM-integrated system: “...The virtual child was engaging, and we had a good conversation! I felt like we made a connection, ...”).

Regarding attitude, we found credible support that people rated the LLM-integrated agent more positively than the rule-based agent (H3). Participants also directly compared the two systems in their comments, with three participants explicitly indicating that the LLM-integrated interaction was better (e.g., “it went better than the first one [i.e., the rule-based]...” and “Definitely better than during the 1st chat [i.e., the rule-based]...”).

Looking beyond the results obtained from hypothesis testing, we observed that participants had a more favourable overall experience with the LLM-integrated agent. Figure 5.5 shows that the LLM-integrated agent scored above the 75th percentile on constructs related to social presence and emotional intelligence. The agent’s emotional presence was also supported by our thematic analysis (e.g., for the LLM-integrated system: “...the virtual child was attentive and behaved more like a human and there were emotions involved.”). Also, participants described the LLM-integrated interaction as more emotionally engaging (e.g., “The interaction was somewhat emotional and eye-opening, reinforcing the importance of empathy, active listening, and proper intervention...”). This aligns with LLM’s capability to generate emotional dialogues [233, 234, 264]. The results of the thematic analysis of participants’ qualitative feedback revealed additional insights into our findings. This reinforces the conclusion that the LLM-integrated system provided a more engaging and believable experience for participants.

On the other hand, the short-ASAQ percentile rating falls within the lower 25% for the LLM-integrated system and 5% for the rule-based system (Figure 5.5). In contrast to the 29 agents in the ASAQ Representative Set 2024, which are primarily assistant agents, our agent is designed as a training simulation. Therefore, lower ratings in some cognitive constructs,

such as performance, attentiveness, and appearance suitability, do not necessarily indicate that the simulation is ineffective. Rather, they may reflect how participants experienced a virtual child in distress, who sometimes “doesn’t know” what to do and needs guidance during the conversation—something a counsellor is expected to provide, not a statement about the system. Response time may have affected attentiveness and performance, as 16 participants noted delays with both agents. However, this is not necessarily a limitation, as real children may also take time to respond. Nevertheless, for training purposes, faster replies might make the experience more engaging. Regarding appearance suitability, one possible explanation is the lack of embodiment, which provides fewer cues during interaction and results in a lack of visible appearance. To gain deeper insights beyond the percentile comparison, the virtual child simulation could be compared more specifically to other chat agents used in social situation simulations of non-experts.

Regarding the experiment’s limitations, the LLM agent had noticeable response delays due to server constraints. Although we matched the response timing in the rule-based agent, participants still reported both systems as slow. For the same reason, we used a lightweight model (LLaMA 3.2) instead of the intended LLaMA 3. Initial tests with the larger model suggest improved comprehension and response quality, which could further enhance system performance. Additionally, interactions in our experiment were limited to single 15-minute sessions. Extending session durations or incorporating repeated interactions might reveal more complex dynamics or improved learning outcomes [191]. Future research could also isolate specific components (e.g., NLU or the bypass mechanism) to assess their individual contributions to learning outcomes and trainee experience. This could provide deeper insights into how the system is perceived and its overall effectiveness.

5.6 STUDY 3: EXPERT INTERVIEWS

The goal of the interviews was to explore experts’ perspectives on hybrid and complete LLM-based training simulations in sensitive contexts, such as child helpline training. We focused on the added value and potential risks of integrating LLMs with an explicit cognitive model. Accordingly, we conducted semi-structured interviews with three helpline experts from 113 Suicide Prevention helpline and De Kindertelefoon. Each interview lasted approximately one hour. This study was approved by the university’s human research ethics committee (HREC reference number: 6218).

5.6.1 PROCEDURE

After signing an informed consent form, experts interacted with the hybrid training system for 15 minutes. We instructed them to explore the system freely, and the interaction could be restarted if it ended before the allocated time. Experts interacted with the hybrid BDI-LLM system described here, and integrated with the pedagogical feedback described in Chapter 3. Following the interaction, experts participated in a semi-structured interview about their experience with the system, focusing on realism, training value, feedback, and controllability. Afterwards, experts were presented with two dialogue scripts. One was generated by the hybrid BDI-LLM system, and the other by an LLM-based training system using GPT-4o, which was prompted to role-play as a child contacting a helpline for five-phase model training, without explicit internal modelling. We then asked the experts

to reflect on the suitability of both dialogues for training purposes, including perceived strengths, weaknesses, potential concerns, and the added value of explicitly modelling the mind in LLM-based simulations.

5.6.2 DATA ANALYSIS

During the interviews, we took detailed notes of the experts replies. To increase reliability, two reviewers analysed these notes independently of each other and identified concerns raised by experts. Afterwards, reviewers discussed their findings and established a final list of concerns. The prompt, the two dialogue scripts, and initial questions guiding the interviews are included in Appendix E.

5.6.3 RESULTS AND DISCUSSION

Based on the interviews, we identified five relevant concerns and opportunities:

Experience With the Hybrid System. Experts generally described the interaction as useful, engaging, and informative for training, and noted that the child understood most things. However, the experience was sometimes perceived as frustrating or unnatural, especially when the child behaved rigidly or misunderstood inputs. While some frustration was seen as realistic and pedagogically useful for creating teachable moments, they mentioned that rigid model structures could reduce the sense of a natural conversation. At the same time, one expert noted that children can do and say strange things, which partly justified such behaviour.

Role of Feedback. Experts valued the feedback mechanisms. They explicitly mentioned five feedback types they liked: undoing the last step, getting advice on the next message, current child situation, current phase progress, and the overall feedback at the end. They emphasised that feedback is essential in such training systems. They also noted that it was beneficial for feedback to be trainee-controlled, i.e., the trainee chooses when and what type of feedback to receive. When asked whether this trainee-controlled approach was better than automatic feedback, experts suggested that new trainees may benefit more from automated guidance, whereas experienced trainees may prefer learner-controlled feedback. Either way, they indicated that both approaches could be useful and that the choice might be left to the trainee.

Teachable Moments and Controlled Challenge. Experts supported the importance of modelling teachable moments in the simulation. They indicated that such moments were clearly modelled in the hybrid system but were not clearly present in the scenario of the LLM-based system. In the hybrid system, the child pursued a challenging goal that created hurdles for the trainee. If the trainee did not deal with the challenge correctly, the child left the conversation. Experts considered these teaching moments informative but found the abrupt leaving of the conversation overly strict. Instead, they suggested that the simulation should show gradual escalation rather than abrupt endings (e.g., warning signals before leaving), which would clearly indicate dissatisfaction. This would give the trainee a chance to recover and steer the conversation back on track. Furthermore, such mistakes should also be clearly reflected in the feedback provided to trainees.

Added Value of an Explicit Cognitive Model. Experts noted that the LLM-based system dialogue expressed the child's desires and concerns more explicitly and with richer context. They mentioned that this made it easier to follow, but sometimes produced uncanny

effects due to overly long and elaborate responses and reduced the controllability of the scenario. In contrast, the hybrid system responses were shorter and more restrained, which some experts perceived as less natural. When comparing the two dialogue scripts, experts highlighted that virtual child simulated by the LLM-based system opened up quickly from the start, whereas the virtual child in the hybrid system was more reluctant and required trust to be established before opening up further. They noted that building trust is an important counselling skill, and that counsellors often find it difficult when someone does not want to go along, requiring more effort to establish trust. Experts also appreciated the feedback explaining the child's thinking, which was based on the BDI model.

Variability in Challenges Is More Important Than in Language. Experts found variability to be useful. The hybrid system was perceived as novel during the first interaction, but became more repetitive when interacting with the same scenario again. In contrast, the LLM-based system could provide more variability and more elaborate replies. Experts agreed that variability in wording alone (i.e., different language for the same behaviour) does not meaningfully improve learning. Instead, what matters is variability in hurdles, challenges, settings, and reactions within the conversation.

5.7 OVERALL DISCUSSION AND CONCLUSIONS

This paper presents a hybrid conversational agent that integrates LLM capabilities into a BDI-based rule system designed for child helpline training. First, we found results that support the non-inferiority of LLM-generated dialogues compared to those generated by humans. Then, through both qualitative and quantitative evaluation, we showed that integrating LLM components for NLU, NLG, and bypass improved believability and trainee attitudes, though the support was weaker for engagement. Also, experts highlighted the importance of teachable moments, structured feedback, and explicit modelling of the child's internal states when using LLMs for simulation-based training.

Several directions for future work could be considered. One direction is to generalise this approach to other systems that train trainees to change individual beliefs and states in conversations, e.g., training social workers to persuade a virtual agent to get vaccinated, or to de-escalate an aggressive agent. This setup reduces the effort required to model conversation content and responses, particularly when crafting training data or generating response variations. However, it still requires designing the rule-based BDI model to capture the pedagogical constraints and structure of conversations, as these are closely tied to the simulation's learning objectives. Furthermore, LLMs could be used to generate belief, desire, and intention contents for a scenario [19].

LLMs controlling or directly influencing the BDI model can be challenging. LLMs are not explicitly trained to convert inputs into structured cognitive states within a numerical or symbolic framework. As a result, biases inherent in LLMs pose challenges when simulating a BDI conversational model. These include formatting and token biases [185, 236], as well as issues from unconventional prompts (e.g., prompts including negation [340]). Therefore, to simulate virtual agents using a BDI framework, LLMs should be paired with a structured BDI system for the LLM to interact with, rather than replace. For example, Retrieval-Augmented Generation [228] could be used to dynamically handle BDI-related content, e.g., retrieving relevant beliefs, desires, and rules during interaction.

One limitation of the second study is that we did not explicitly analyse the extent to

which interactions were redirected to the bypass component. If, for example, all interactions were handled through the bypass, improvements in believability could have been partly due to unconstrained LLM responses. However, results from the first study indicate that the intent recognition component performed reliably, suggesting that the system did not predominantly rely on the bypass component. This observation is further supported by the third study, in which the bypass component was triggered only five times out of 78 inputs across three expert interactions. This indicates that the majority of the decision-making is still controlled by the BDI model.

Although we focused on integrating LLMs into the agent simulator component, the ARTES architecture [8] highlights other potential integration points. For example, LLMs could be used to generate feedback based on trainees' interactions, enabling deeper skill acquisition [63, 90, 402]. In our case, the BDI system could support feedback generation by supplying the LLM with structured inputs, such as the child's internal state, conversational logs, and learning materials. Furthermore, an LLM could simulate a senior counsellor, offering additional information and answering trainees' queries during interactions.

To conclude, the hybrid approach maintained pedagogical control over the conversational structure, yet offered a better balance between realism and control. This balance opens up new opportunities for scalable, realistic training in sensitive interaction domains, while still aligning with specific learning objectives.

6

CONCLUSION

6.1 MAIN CONCLUSIONS

This thesis studies how to train lay counsellors at children's helplines using an agent-based simulation tool. To this end, we first examined the broader field of social skills training systems to identify their key components (Chapter 2). Building on these components, we empirically investigated how simulation, feedback, and reflection contribute to learning outcomes by studying their incremental effects within a chatbot-based training environment (Chapter 3). We then examined whether explicitly modelling human values within an agent's internal states improves trainees' understanding of the simulated interaction and perceived alignment with the virtual child (Chapter 4). Finally, we investigated how integrating a Large Language Model with a rule-based BDI model affects the perceived believability, consistency, and suitability of open-ended interactions for training purposes (Chapter 5). Through this work, we aimed to answer the following overall research question:

Main Research Question

In what ways can interactive simulation-based learning tools enhance the training of counsellors in children's helplines?

From this main research question, we built the thesis through one initial exploratory research question followed by three hypotheses. The conclusions for each are presented below.

Subquestion 1

What are the key components of simulation-based social skills training systems, and how can they be structured into a comprehensive architecture?

The research identifies two key components of simulation-based social skills training systems: the agent simulation and the educational components. Based on an analysis of 43 training systems and 8 existing architectures, we propose the ARTES architecture,

which structures these components and clarifies their roles. For training systems aimed at changing thoughts and beliefs, two components are particularly central: ‘Thinking’ within the simulation and the ‘Tutor’ within the education module. ‘Thinking’ represents the agent’s decision-making process and model, including dialogues, states, emotions, and backstory. These components can be represented in various ways, such as direct question-answer pairs, cognitive models (e.g., BDI), or data-driven approaches. The ‘Tutor’ handles the pedagogy of the training, either by controlling the task or providing feedback and guidance.

Mapping the ARTES components to four learning theories showed how the components support different learning objectives. Behaviourism emphasises introducing teachable moments with specific triggers to reinforce trainee behaviour; cognitivism focuses on teaching and explaining situations; Social Cognitive Theory involves learning from diverse situations; and constructivism supports exploration and facilitating the learning environment. ARTES also demonstrated coverage and generalisability when mapped to the analysed systems and architectures. We also identified nine future research directions: knowledge transfer, realism-education balance, explainability, ethics, LLMs, crowdsourcing solutions, multi-agent scenarios, research methods, and procedurally generated scenarios.

Hypothesis 1

Simulation with feedback and reflection in social skills training systems improves learning.

The research provides credible support for the first hypothesis. We tested the hypothesis through a longitudinal randomised controlled trial ($N = 346$) comparing additive configurations of simulation, feedback, and reflection. The results demonstrate an overall increase in learning with the subsequent addition of components. When compared stepwise, adding simulation improved learning relative to the control condition, which involved only watching an educational video. Adding feedback to the simulation further improved learning compared to the simulation alone. However, adding reflection tasks on top of this dampened improvements in learning when compared to simulation with feedback, for which there is weak support. In a similar pattern, secondary measures of perceived usefulness, user experience, and usability increased as components were added. Overall, the results show that simulation systems, particularly those with feedback, improved learning outcomes, whereas adding reflection improved reflective capability but appears to slow the rate of performance improvements relative to simulation with feedback.

Hypothesis 2

Integrating human values in a simulation improves the learning experience.

The results of this thesis provide credible support for the second hypothesis. We conducted a between-subjects experiment ($N = 193$) comparing a base BDI agent, a value-integrated agent, and a value-integrated agent with explanatory feedback. Participants interacting with agents that explicitly modelled values showed higher value recognition and situational understanding compared to the base condition. Measures of perceived

alignment opportunities increased in the value-integrated condition, while adding feedback showed no clear additional effect on this measure. A similar pattern was observed for the remaining measures: perceived usefulness, believability, and overall experience. Furthermore, participants reported values they perceived in the virtual child, which matched the values we intentionally designed. Overall, the results confirm the hypothesis that incorporating values within a training scenario, in both the simulation and the feedback, improved value recognition, situational awareness of the values, believability, perceived opportunity to align with the values, and perceived usefulness.

Hypothesis 3

Integrating cognitive models and LLMs in a simulation improves the learning experience.

The thesis provides credible support for the third hypothesis. Evidence was obtained through three complementary studies evaluating a hybrid BDI-LLM agent. (1) In a script-based non-inferiority study, LLM-generated responses and intent recognition were not inferior to human-generated content. (2) In a within-subjects experiment ($N = 37$), participants rated the hybrid agent as more believable, reported a more positive attitude toward it, and preferred it over the rule-based agent. This preference was also reflected in qualitative responses. (3) Expert interviews further supported the suitability of integrating LLMs within a structured cognitive model for training scenarios, as such models could help trainees understand why a child might exhibit specific behaviours. Overall, the empirical evidence indicates that integrating LLMs with a cognitive model was positively perceived by both trainees and experts.

6.2 LIMITATIONS

There are several limitations to consider when interpreting the findings of this thesis. Although the initial design and evaluation of Lilobot involved counsellors from the Dutch children's helpline, the present thesis did not include an empirical evaluation with actual volunteers or trained counsellors. Instead, to match the target population, we recruited participants on Prolific who met the Dutch Child helpline entry criteria (i.e., holding at least a Bachelor's degree). Prolific workers generally provide higher-quality data, which has been shown both in comparison to other crowd-sourcing platforms and to undergraduate samples [106]. To avoid self-interest and loss aversion biases, we clarified that participants' payment would not depend on their performance or interactions [107]. However, crowd-sourcing workers' motivation goes beyond just payment. Their motivation is also influenced by factors relevant to the experiments here: enjoyment, skill development [404], and contributions to a social cause [69]. When looking at the results, participants showed skill improvement; for example, they improved over time (H1) and identified the intended modelled values (H2). On the other hand, paid participants might have felt more obligated to complete all learning tasks, while a helpline counsellor might skip learning tasks they deem unnecessary. This implies a difference in motivational structure: crowd-sourcing workers' motivation may stem from the factors described above, while volunteers are typically intrinsically motivated [263], e.g., by commitment to the helpline and its mission.

The training examined in this research focused on teaching trainees how to change virtual children's thoughts and beliefs. We achieved this by integrating an internal model that guides the conversation and bases feedback on the model. This approach showed generalisability across two communication protocols: the five-phase model and motivational interviewing. It remains to be explored whether this approach also generalises to other types of chats received by helplines, such as pranks [391], repeat contacts that require managing dependency, boundaries, and continuity of support [306], or very sensitive or high-risk situations (e.g., exploitation and trafficking [382]). In such cases, the conversational goal is not to help reshape the child's internal states by following a communication protocol. Rather, trainees should learn to recognise signals and act appropriately, while also managing their own emotional state. Such sensitive situations may lead to emotional exhaustion and burnout among helpline counsellors [94, 232], and task performance training can help manage such stress [72, 225]. Also, self-reflection might focus on broader categories in this training, involving not only learning-oriented reflection but also feelings and personal beliefs [371].

Although we simulated interactions with a virtual child in this research, we did not involve children directly, for example, by co-creating the virtual child's personas or conversational patterns. Including children directly could have resulted in more realistic behaviours and more relatable personas. Nevertheless, children were intentionally not included for ethical reasons, such as exposing minors to sensitive topics. The primary learning objective of the training was to support trainees in practising the five-phase model, rather than to maximise conversational realism. To support this objective, the personas and scenarios were developed with the help of experts, which contributes to their validity. Thus, although child involvement in the design might have further improved ecological validity, we expect it to be unlikely to have altered the main learning effects observed.

Regarding evaluation, we assessed participants' performance and task knowledge (H1), and self-reported measures (H1, H2, H3). Although we focused on the intended target group, we did not examine long-term retention or the transfer of knowledge to real-world contexts, as has been shown in other simulation-based training domains (e.g., diagnostic simulations [127] or surgical simulations [337]). Knowledge transfer is an important consideration for educational systems, as successful training should support learners in applying knowledge beyond controlled environments. Prior research suggests that short-term retrieval practice in educational settings can enhance long-term retention [316, 352]. Future work should evaluate training impact in additional settings, such as role-playing exercises or supervised training sessions, to assess the training system's ability to improve trainee performance in realistic scenarios, rather than unsupervised interaction with real children. Evaluating performance with real children raises ethical and practical concerns, e.g., in experimental manipulation, whereas role-playing and supervised sessions allow realistic yet controlled assessment of counselling performance. Thus, experienced trainers or supervisors could assess trainee performance using established rubrics, enabling evaluation of skill transfer while safeguarding children's well-being.

For the statistical analysis, we used Bayesian methods because they estimate how probable the hypotheses are given the data, allowing results to be interpreted beyond a binary significant/non-significant decision. A limitation is that readers who are more familiar with frequentist testing may find this reporting style less familiar. Another

limitation regarding analysis is that sample size estimation was based on a frequentist power analysis as a simpler approximation, rather than on a simulation-based Bayesian sample size analysis. However, frequentist and Bayesian approaches often yield similar sample size estimates under comparable assumptions [175]. A further limitation is that we assumed a medium effect size as a conventional benchmark for the power analysis. This may have reduced our ability to detect small effects with high certainty. However, it remains debatable whether such small effects would be practically relevant or noticeable.

Finally, the contributions of this thesis primarily focus on the perspectives of trainers and trainees, that is, whether the learning goals were modelled appropriately and whether participants improved. However, we did not investigate how such a training system should be integrated into educational practice or how educators might adapt it within a broader curriculum. This is important as educational design research emphasises alignment between learning materials, instructional methods, and assessment tools [11, 187, 246, 252]. While the empirical insights presented in this thesis, for example on the effects of feedback on learning outcomes, may inform future integration efforts, studying the scalability and costs of deploying and maintaining such systems remains an important direction for future work.

6.3 FUTURE WORK

We identify four main directions for future work: the agent simulation, feedback and guidance, the overall training structure and goals, and deployment and research.

Agent’s Simulation: Although we integrated LLMs into the NLU and NLG components, decision-making remains fully dependent on the rule-based BDI model. As discussed earlier, internal models are important to the training setting explored in this thesis. Future work could investigate different ways of integrating internal states into LLM-based agents, and assess how they affect realism, controllability, and learning outcomes. A possible approach is to combine LLMs with a BDI model to replace part of the rule-based decision-making process. This could increase realism while reducing the effort required to manually model all decision rules. One example of such integration is the use of LLMs to recognise emotions from dialogue inputs in a zero-shot prediction and inject them as beliefs into a BDI agent [304]. This allows the agent’s behaviour to adapt emotionally without explicitly relying on rule-based emotion recognition models.

A promising future direction is to examine how different approaches to utterance generation affect the training experience. Rather than relying on manually authored dialogue, language models could be trained or fine-tuned on helpline or counselling data to increase response variability and fluency (similar to the approach in [283]). However, it is important to consider the risk that an LLM may reproduce utterances from real conversations. One possible mitigation strategy is to generate synthetic data based on real conversations and train models on this synthetic data to add a layer of anonymisation, or to rely on foundation models that are not trained on helpline-specific data.

Feedback and Guidance: In the current training setup, feedback is rule-based and tied to the modelled situations and internal states, which constrains how flexibly trainee actions and their consequences can be explained. LLMs can also be used to generate feedback based on learning objectives and the trainee’s interaction with a virtual help-seeker. Several studies have explored LLM-generated feedback in simulation-based training systems, for

example, to train medical students with virtual patients [63] or to support novice peer counsellors [75]. This could be extended by explicitly incorporating the agent's internal states into the feedback. In this way, the LLM would not only provide feedback on trainee performance, but also explain the help-seeker's underlying beliefs, goals, and how the trainee's actions affected them.

Another avenue is to support knowledge building through guidance during training. Such guidance could reduce cognitive load, support trainees, and promote skill acquisition. For example, the system could introduce a virtual senior counsellor who guides the trainee. This virtual mentor could provide feedback, explain concepts, motivate trainees, or assist with difficult moments during the interaction. Additional support could be provided by showing parallel or similar example dialogues that resemble the current interaction, allowing trainees to draw inspiration from previous successful cases [326]. Furthermore, trainees' knowledge development could be supported through instructional principles that structure tasks and learning progression, for example, via scaffolding mechanisms [330].

Training Structure and Goals: The proposed approach in this thesis could be generalised to other communication protocols and to different helplines by applying the same internal-states modelling approach. In addition, the approach could be adapted to train counsellors for other sensitive scenarios, such as exploitation and trafficking. This could also be extended to other target groups, for example, teachers learning to recognise and respond to emotional or behavioural triggers in students. Some of these applications may require additional design considerations, such as an agent embodiment. Related work has already explored simulation-based training for laypeople to support conversations about health behaviours, e.g., vaccination, with family members and friends [283]. Beyond health contexts, a similar approach could be applied in non-health domains, for example, to train customer service representatives to handle conflicts in chat-based interactions.

The training system in this thesis draws primarily on behaviourism, cognitivism, and constructivism. Other learning theories could also be explored. For instance, social cognitive theory could be incorporated by allowing trainees to observe and compare multiple perspectives in multi-agent scenarios (e.g., [12]). From a humanistic perspective, training could focus more explicitly on supporting self-reflection and self-actualisation of trainees as developing counsellors, e.g., through an LLM-based conversational agent that supports reflective thinking [397].

Furthermore, the training system could be made more modular by developing a user-friendly authoring tool. Such a tool would allow educators, without technical expertise, to create and adapt BDI-based scenarios tailored to specific learning objectives or trainee profiles. This could be extended by using AI to initialise agent personas from values, beliefs, and internal states while maintaining behavioural and narrative coherence, for example, using frameworks such as *DiriGent* [399].

Deployment and Research: Although we evaluated a stand-alone training simulation, future work could examine how such systems can be integrated into existing curricula, instructional methods, and organisational training structures. This includes understanding how simulations can complement role-playing, supervision, and theoretical instruction. Also, identifying practical requirements for large-scale deployment, such as time investment, trainer involvement, and technical infrastructure. In addition, as discussed in the limitations, future studies should examine the transfer of learning outcomes to real-world

counselling practice.

Another important research direction concerns trainee motivation and engagement. For example, reinforcement learning techniques could be used to build up competencies [9], such as by adapting feedback and guidance strategies over time. Early stages of training might prioritise motivational feedback, while later stages could focus more on performance improvement. Finally, future research could investigate how the values of trainees, trainers, or system designers influence the training process, or vice versa. This includes examining whether and how training shapes trainee values over time, and to what extent such value shifts are desirable or ethically acceptable. For instance, by analysing their reflections or conversations to estimate shifts in value preferences, in line with approaches that infer values from choices and motivations [231].

6.4 CONTRIBUTIONS

SCIENTIFIC

Modelling of Internal States: This work contributes a model of a child's internal state for use in agent-based systems and demonstrates how it can be implemented to support learning in simulation-based training environments. Existing research on simulation-based helpline training focuses primarily on observable interaction behaviour, with less attention paid to the motivations underlying a child's actions. The thesis shows that explicit cognitive modelling of a help-seeker's internal state, in this case through a BDI model, allows internal states to be represented and controlled as part of training aimed at changing help-seekers' beliefs. This cognitive model is further extended to integrate underlying values that influence a child's thinking and decision-making. This modelling enables the intentional design of diverse profiles in interactive agents. In addition, the thesis demonstrates how cognitive models can be combined with LLMs to improve language understanding, response diversity, and perceived realism, while balancing control, transparency, and interpretability.

Design of Training Tools: At a general level, we propose the ARTES architecture, a modular architecture for agent-based training systems. Existing conceptual architectures of training systems often prioritise either virtual human simulation (e.g., [155]) or tutoring processes (e.g., [14]), but rarely make explicit how these two parts should be combined within a single architecture. ARTES addresses this by synthesising insights from agent-based simulation, intelligent tutoring systems, and learning theories. The architecture provides researchers with a structured way to analyse, compare, and understand training systems. By making system components, interaction characteristics, learning-theory focus and design trade-offs explicit, ARTES supports future research in the field.

Beyond the conceptual contribution of ARTES, this thesis also contributes a concrete training system that integrates simulation with feedback and reflection mechanisms. Such an environment can be extended or adapted to conduct experiments, either to examine how learning can be improved or to explore how people use these behaviours under specific controlled cues. The work shows how feedback can be generated based on internal cognitive states, such as beliefs and values, and how this feedback relates to different feedback levels (e.g., task-level and process-level) and timings (during versus after interaction). Furthermore, the design of reflection activities and the overall structure of training sessions are grounded in learning theories. Overall, these components provide a blueprint and an

example to follow for adding feedback to similar training systems.

Empirical Evaluation: This thesis provides empirical evidence on the effectiveness of agent-based simulation for training lay counsellors, showing how pedagogical structure, cognitive modelling, and interaction design influence learning and user experience. Through controlled experiments and a large randomised controlled trial, the work clarifies the contribution of each element to learning, rather than treating training systems as a single package. When mapping the empirical evaluation to Kirkpatrick's Four Levels of Training Evaluation [202], it primarily addresses Level 1 (Reaction; how trainees perceive the training) and Level 2 (Learning; acquired knowledge, skills, and understanding of the intervention). Evaluating higher levels—Level 3 (Behaviour; the transfer of learned skills to real counselling practice) and Level 4 (Results; the impact of training on organisational or societal outcomes)—remains unexplored for simulation-based training systems. This can be attributed to the complexity of measuring real-world behavioural impact, as well as the substantial time and monetary costs involved in conducting such evaluations for virtual agents in social skills training [48].

The thesis empirically examines trainees' experience with value-based cognitive models in training. The results show that participants preferred the integration of values and perceived it as better aligned with the intended training objectives, such as recognising and responding to value-sensitive cues. The hybrid LLM-BDI system was also perceived as more realistic and effective. The thesis also identifies future challenges and research directions for the field, such as ethical and scalable agent-based training systems.

We also make the experimental structure, training scenarios, system implementations, evaluation instruments, and analysis procedures explicit. By adopting open research practices, including preregistration and sharing of materials and code, this work enables others to reproduce, replicate, compare, and extend the reported findings, thereby supporting cumulative research in the field.

SOCIETAL

Trainees: This research supports trainees at helplines, such as child helpline counsellors, in addressing the challenge of acquiring and maintaining communication skills for complex and emotionally sensitive conversations. Applying communication protocols adequately typically requires repeated practice, which is difficult to provide consistently through traditional training alone. The simulation-based training tools developed in this thesis offer trainees a safe and readily available environment for such practice. In addition, these simulated experiences help trainees recognise cues from help-seekers and build knowledge through experience. The empirical evaluation showed improvements in proximal outcomes, including knowledge test scores and performance within the simulated training environment. Although this does not provide direct evidence of transfer to real helpline conversations, it does suggest that the training supports the relevant skills.

Organisations and Help-seekers: Helplines face large volumes of contacts and pressure to recruit and train new volunteers. Although the training systems proposed in this thesis are not positioned as an alternative to traditional training, simulation-based training can help alleviate the burden on training resources [58, 186]. This may allow organisations to better prepare new trainees without relying excessively on trainers' time for day-to-day training. As a result, trainers can focus their efforts on supporting trainees

in more complex and nuanced situations. In addition to efficiency, the BDI model allows educators to vary help-seeker profiles to create diverse training scenarios. These scenarios are not necessarily tied to learning the five-phase model alone, but can be designed to support different learning objectives. This allows organisations to expose trainees to a wider range of situations, perspectives, and training goals. In turn, this could also reduce the risk that trainees encounter unfamiliar or challenging situations for the first time during real conversations with vulnerable help-seekers. For help-seekers (i.e., helpline clients), more efficient training could translate into safer, more appropriate, and more empathetic support [34, 183]. Trained counsellors who have practised handling emotional and value-sensitive conversations are better equipped to respond appropriately, empathetically, and in line with communication protocols.

Developers, Designers, and Educators: For system designers and developers, ARTES offers a conceptual foundation that supports building training systems from scratch while making design decisions explicit and comparable. For educators, the architecture highlights the role of learning theories in training design. For all three groups, the thesis clarifies how elements such as simulation, feedback, and reflection can be systematically embedded into interactive training tools.

We also present a BDI-based computational model for changing a simulated help-seeker's beliefs through interaction. Based on that, the thesis introduces a mechanism for generating feedback based on the agent's internal cognitive state. These show how internal models can be used not only to drive agent behaviour, but also to support explanation and instructional feedback across different simulation-based training systems.

Finally, the materials produced in this thesis offer practical value for reuse, comparison, and further research. By providing concrete scenarios, system designs, system code, evaluation instruments, and empirical results, the work enables developers, designers, and educators to compare alternative system designs, replicate and extend training systems, and adapt existing materials to different domains, populations, or learning objectives.

6.5 ETHICAL REFLECTION AND IMPLICATIONS

The development and evaluation of a training application for sensitive contexts, such as helplines, should adhere to ethical standards and integrity. This section reflects on the ethical considerations involved in conducting this research and the broader implications of the research outcomes.

Reproducibility and Replicability: To ensure the transparency and reliability of our findings, we adhered to open science principles throughout this research. All experimental data, analysis code, system source code, and questionnaires (if possible) have been made publicly available via the 4TU Research Data Repository and GitHub. By sharing these materials, we aim to minimise issues related to replication and allow other researchers to verify or build upon our work. Furthermore, all studies presented in this thesis were preregistered on the Open Science Framework (OSF) before data collection. This is to reduce the risk of selective reporting and improve reproducibility. We also conducted reliability checks on qualitative coding, such as using multiple coders for thematic analysis and validating LLMs' scoring against human evaluators.

Participants' Involvement: All experiments received approval from the Human Research Ethics Committee (HREC) of the Delft University of Technology. To ensure

participant privacy, all data were manually anonymised, stored securely on university servers, and only anonymised datasets were published or shared outside the research team. Furthermore, we informed participants of the study's nature and potential risks before asking them to sign a consent form. Only those who signed the form and agreed were included in the experiments. We recruited participants online via Prolific and compensated them in accordance with Prolific's ethical guidelines [309]. For the single sessions studies (Chapters 4 and 5), compensation was aligned with the platform's minimum wage guideline (at least £6 per hour). For the longitudinal study (Chapter 3), participants received the same rate for the first four sessions but were offered a higher rate (£10 per hour) for the final session. This increase in the compensation structure was intended to motivate retention over the two-week study period and to fairly compensate participants for their continued participation. It is worth noting that although this minimum rate aligns with Prolific guidelines, minimum wages vary globally. Consequently, what constitutes fair compensation may differ by participant location, which represents a general ethical limitation of crowdsourcing platform research.

Improper Training and Misuse: Our simulations relied on a finite number of scenarios, personas, and communication protocols. For example, Chapter 3 introduced twelve scenarios primarily focused on bullying, and Chapter 4 explicitly modelled two children with specific value profiles. Modelling only limited behaviours risks reinforcing stereotypes about how children think or act when contacting helplines. If trainees are exposed solely to these specific representations, they may develop misaligned expectations about real-world interactions. Training, therefore, should include diverse perspectives and behavioural patterns to promote deep understanding and transfer [351], and to reduce the risk of stereotyping or overgeneralising specific values, behaviours, or actions [303, 367].

In the thesis, the core objective of our training system is to teach trainee counsellors to influence a help seeker's internal beliefs and thoughts. While our goal is training to support children in finding safer solutions, this still poses a risk of misuse. The underlying mechanism of belief change could, in theory, be modelled to train people to manipulate others or engage in non-prosocial behaviours. For example, how to manipulate people's beliefs into complying with actions that primarily serve the manipulator's interests. Trainees need to be aware of the ethical boundaries of training simulations to ensure they are used to support the help seeker's autonomy rather than undermine it. This concern directly relates to regulations such as the EU AI Act [115], which restricts AI systems that intentionally manipulate individuals' behaviour or exploit vulnerabilities.

AI Integration in Education: The integration of AI in education, particularly LLMs, introduces ethical challenges. UNESCO's report on AI and the future of education discusses several concerns [372], including transparency, privacy, explainability, and controllability, which are also general concerns across other application domains. The report highlights the interactive nature of AI agents compared to passive tools, noting that such systems may influence belief formation rather than merely supporting content delivery. It also highlights risks related to over-reliance on AI and cognitive offloading. This may undermine deeper learning and increase the likelihood that people adopt AI-generated output without adequate reflection. Furthermore, the report raises concerns about inequality, as the infrastructure and access required to use such systems for education are not available to at least one-third of the world's population. In their book on the ethics of AI in education,

Holmes and Porayska-Pomsta [170] also discussed the ethics of nudging and the risk of declining human agency. Both sources additionally emphasise risks related to data governance and the ethical deployment of AI in contexts involving vulnerable learners. Together, these concerns highlight the need for ethically grounded, human-centred educational AI systems. In this thesis, these concerns are addressed by grounding AI behaviour in explicit, interpretable agent models and pedagogical structures. For example, modelling reasoning in a rule-based BDI preserves transparency and controllability, while supporting realistic, explainable and ethically responsible learning interactions.

Beyond social and ethical concerns, the growing use of AI in education also raises environmental concerns. Recent work highlights that LLMs incur substantial environmental costs throughout their lifecycles, including energy consumption, water use, and greenhouse gas emissions [238]. These impacts arise not only during model training, but also during deployment and everyday use. Going forward, it is important to consider environmental sustainability alongside ethical and pedagogical concerns when integrating AI into educational systems, especially when deployed at scale.

6.6 FINAL REMARKS

Children's helplines rely on counsellors who can communicate with empathy, structure conversations effectively, and support children in finding solutions. Developing these skills requires repeated practice, which could be costly in traditional role-playing settings. At the same time, there is a high demand for helplines, which requires training more counsellors, further amplifying these costs. Interactive simulation-based tools could address these challenges by offering an accessible, private, and safe training environment. Such an environment allows trainees to interact with a virtual child to apply their knowledge, while receiving guidance to build and refine their skills. This thesis aimed to explore how simulation-based learning tools can better support the training of children's helpline counsellors. This resulted in (a) an architectural framework for social skills training systems, (b) a randomised controlled trial demonstrating that combining simulation with pedagogical guidance improves learning outcomes, including knowledge and performance, (c) empirical evidence that modelling beliefs and values increases interpretability, realism, and alignment with counselling goals, and (d) empirical evidence that integrating large language models with cognitive models enables natural yet controllable interactions. Together, these findings show how interpretable, value-aware AI simulations, when combined with pedagogical guidance and feedback, can support effective and responsible social skills training. This highlights the potential of simulation-based training to strengthen and accelerate educational practice in sensitive contexts, reinforcing both professional competence and educational impact.

A

APPENDIX FOR CHAPTER 1

A.1 OVERVIEW OF PREVIOUS WORK ON LILOBOT

Lilobot was developed as a conversational agent to support the training of counsellors at children’s helplines in previous work [145], as part of an MSc thesis [144]. Lilobot simulates a child who reaches out to a helpline via text-based chat. The training objective is to help trainees practise a structured, child-centred conversation strategy used in helplines. Trainees interacted with Lilobot via a chat interface and attempted to guide the conversation according to this strategy.

A.1.1 LILOBOT’S SCENARIO DESIGN

To support Lilobot’s learning objective, one scenario conversation was crafted in collaboration with experts from the Dutch children’s helpline, without the use of LLMs. Lilobot’s scenario focused on bullying at school. This scenario was deliberately configured to include challenging conversational elements. For example, Lilobot could request that the helpline directly contact the school to stop the bullying. This request falls outside the helpline’s scope and was included to guide the trainee in applying the conversation model carefully, rather than simply responding to the child’s first stated wish. The selected persona and bullying context function as an example, while the same BDI-based method (explained later) can be adapted to other personas and scenarios.

A.1.2 THE FIVE-PHASE MODEL

Lilobot was developed to train people to apply communication protocols. In this case, the protocol was the five-phase model [344], a conversational strategy used in children’s helpline services. Table A.1 shows the five-phases in more detail. The trainee was expected to follow these phases during the conversation. If the trainee deviated from these phases, for example, by moving too quickly, showing insufficient empathy, or failing to clarify the child’s situation, this could lead to less desirable conversational outcomes, e.g. the child leaving the conversation.

Table A.1: The five-phase Model steps [344].

Phase Name		Function in the conversation
1	Building rapport	Establish contact with the child and create a safe atmosphere for the conversation.
2	Clarifying the child’s story	Explore the child’s situation, understand what happened, and identify the child’s perspective and concerns.
3	Setting the goal	Clarify what the child wants from the conversation and work toward a realistic and appropriate conversational goal.
4	Working toward the goal	Support the child in identifying possible plans to implement the goal and encourage actions that fits the child’s situation.
5	Rounding off the conversation	Conclude the conversation appropriately, and in a supportive way.

A.1.3 LILOBOT'S THINKING

Lilobot's thinking process is based on a Belief–Desire–Intention (BDI) computational model. Lilobot holds various beliefs about itself, the trainee, and the world, which are affected by the trainee's input during a conversation. Together, the specific beliefs, desires, and intentions defined in this model form Lilobot's persona described above; changing this content would result in a different persona. These beliefs function as parameters on an interval scale from 0 to 1 and can increase or decrease depending on how the trainee responds. For example, whether Lilobot thinks the trainee can be trusted, whether the trainee understands its story, and whether the conversation feels safe.

Lilobot's desires represent the child's goals. For example, Lilobot may want to talk about its problem, want the trainee to solve the problem directly by calling the school, want to work toward a solution, or want to end the conversation. The intention is the desire that is currently active and being pursued by the agent. Lilobot adopts intentions based on the beliefs it currently holds. For example, if the belief that the trainee can be trusted drops below a specific threshold, the child may adopt the desire to end the conversation. Based on its current internal states, Lilobot selects an appropriate action from its knowledge base. In this way, Lilobot's behaviour is chosen through the interaction of its beliefs, desires, and intentions. This structure was important for the later work in this thesis, as it provided an interpretable cognitive basis for the simulation and made it possible to represent not only what the virtual child says, but also why its behaviour changes over time.

To recognise trainee inputs and match them to belief changes, Lilobot used a pre-trained NLU model built with Rasa [44]. When the trainee sends a message, the system first classifies the raw text input into a predefined trainee intent. This recognised trainee intent is then passed to the BDI model for reasoning. The reasoning engine uses a set of predefined rules that map each trainee's intent to changes in Lilobot's belief values and, where relevant, to changes in its active desire and next action. For example, if a trainee's input is recognised as showing empathy, Lilobot's belief that the trainee understands it increases, which can lead Lilobot to share more information about its problem. After updating its beliefs, Lilobot re-evaluates which desire should become active. The active desire then determines the intention Lilobot is currently pursuing, after which Lilobot selects the next response or action. This process is repeated for each trainee's input. In addition, Lilobot can initiate a message if the trainee remains inactive for a short period. In such cases, the agent retrieves a response associated with its current desire, thereby maintaining the conversational flow. A detailed, line-by-line conversational example illustrating how different trainee inputs affect Lilobot's BDI model is described by Grundmann et al. [145].

Lilobot's architecture (Figure A.1) combines a natural language understanding component with a rule-based BDI reasoning engine. The NLU component recognises the trainee's intent, while the reasoning engine applies predefined rules to update Lilobot's internal states and select a response. The selected response is then returned to the trainee through the chat interface.

Although the original implementation was rule-based and therefore limited in dealing with unexpected or highly varied trainee input, it established the core simulation framework for the later studies in this thesis. Thus, it provided the initial interpretable agent structure that was later extended with additional scenarios, educational components, value modelling, and the integration of large language models.

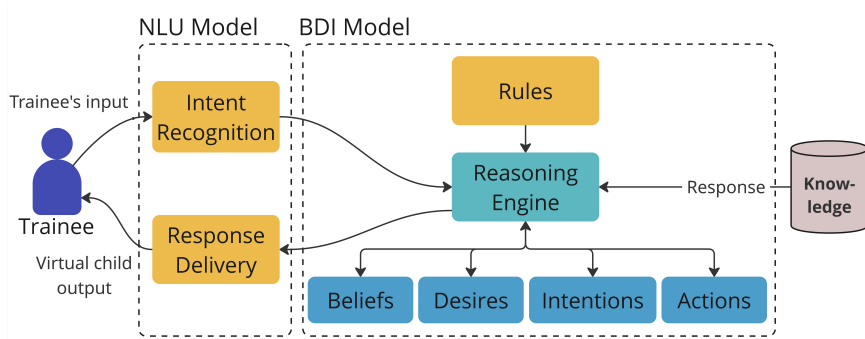


Figure A.1: Architecture of Lilobot [145].

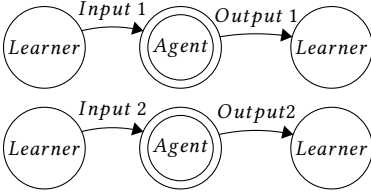
B

B

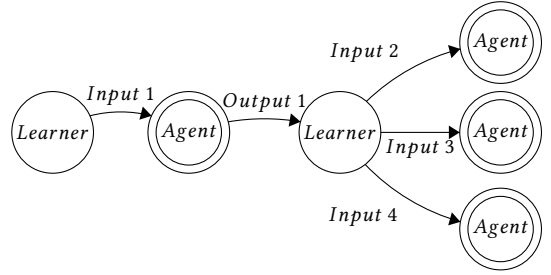
APPENDIX FOR CHAPTER 2

B.1 FURTHER EXPLANATIONS OF CATEGORISATIONS

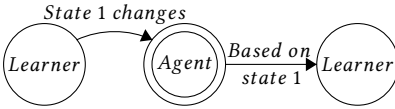
B



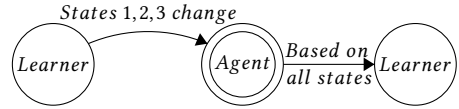
(a) Action decision: The input is directly mapped to output (⚙️2).



(b) Action decision: a scenario is defined (⚙️3).



(c) Agent model: limited states changes (⚙️2).



(d) Agent model: representation of changes (⚙️3).

Figure B.1: State machine diagrams explaining the four categorisation mentioned in Table 2.5. "Learner" and "Agent" refer to their replies.

C

C

APPENDIX FOR CHAPTER 3

C.1 PARTICIPANT DEMOGRAPHICS

Table C.1: Full participant demographics by condition (control, simulation only, simulation and feedback, simulation, feedback and reflection).

Demo-graphic	Control (n=66)	Simulation (n=66)	Simulation and feedback (n=64)	Simulation, feedback, and reflection (n=150)
Age	$M = 31.41, SD = 9.58$ 18–23: 7 (10.6%) 24–29: 28 (42.4%) 30–39: 24 (36.4%) 40–49: 5 (7.6%) 70+: 2 (3%)	$M = 31.50, SD = 9.13$ 18–23: 14 (21.2%) 24–29: 17 (25.8%) 30–39: 26 (39.4%) 40–49: 4 (6.1%) 50–59: 5 (7.6%)	$M = 31.75, SD = 8.59$ 18–23: 13 (20.3%) 24–29: 17 (26.6%) 30–39: 24 (37.5%) 40–49: 6 (9.4%) 50–59: 4 (6.2%)	$M = 31.90, SD = 10.24$ 18–23: 30 (20%) 24–29: 48 (32%) 30–39: 48 (32%) 40–49: 12 (8%) 50–59: 6 (4%) 60–69: 5 (3.3%) 70+: 1 (0.7%)
Gender	Men: 30 (45.5%) Women: 36 (54.5%)	Men: 28 (42.4%) Women: 38 (57.6%)	Men: 32 (50%) Women: 32 (50%)	Men: 62 (41.3%) Women: 81 (54%) Non-binary: 4 (2.7%) Prefer not to say: 3 (2%)
Education	Community/Technical: 3 (4.5%) Bachelor's: 44 (66.7%) Master's: 15 (22.7%) PhD: 4 (6.1%)	Community/Technical: 3 (4.5%) Bachelor's: 49 (74.2%) Master's: 12 (18.2%) PhD: 2 (3%)	Community/Technical: 6 (9.4%) Bachelor's: 41 (64.1%) Master's: 16 (25%) PhD: 1 (1.6%)	Community/Technical: 13 (8.7%) Bachelor's: 99 (66%) Master's: 31 (20.7%) PhD: 7 (4.7%)
Has Children	Yes: 30 (45.5%) No: 34 (51.5%) Prefer not to say: 2 (3%)	Yes: 32 (48.5%) No: 30 (45.5%) Prefer not to say: 4 (6.1%)	Yes: 36 (56.2%) No: 25 (39.1%) Prefer not to say: 3 (4.7%)	Yes: 82 (54.7%) No: 62 (41.3%) Prefer not to say: 6 (4%)
English Fluency	Native: 27 (40.9%) Fluent: 39 (59.1%)	Native: 27 (40.9%) Fluent: 39 (59.1%)	Native: 25 (39.1%) Fluent: 39 (60.9%)	Native: 53 (35.3%) Fluent: 97 (64.7%)
Field of Study	Business: 6 (9.1%) Law: 6 (9.1%) ICT: 10 (15.2%) Engineering: 13 (19.7%) Natural Sciences: 2 (3%) Social Sciences: 2 (3%) Medical Sciences: 3 (4.5%) Humanities: 7 (10.6%) Education: 6 (9.1%) Other: 11 (16.7%)	Business: 10 (15.2%) Law: 2 (3%) ICT: 13 (19.7%) Engineering: 5 (7.6%) Natural Sciences: 6 (9.1%) Social Sciences: 2 (3%) Medical Sciences: 6 (9.1%) Humanities: 1 (1.5%) Education: 8 (12.1%) Other: 10 (15.2%) Prefer not to say: 3 (4.5%)	Business: 12 (18.8%) Law: 4 (6.2%) ICT: 8 (12.5%) Engineering: 5 (7.8%) Natural Sciences: 1 (1.6%) Social Sciences: 2 (3.1%) Medical Sciences: 11 (17.2%) Humanities: 3 (4.7%) Education: 9 (14.1%) Other: 8 (12.5%) Prefer not to say: 1 (1.6%)	Business: 20 (13.3%) Law: 8 (5.3%) ICT: 36 (24%) Engineering: 17 (11.3%) Natural Sciences: 7 (4.7%) Social Sciences: 8 (5.3%) Medical Sciences: 11 (7.3%) Humanities: 7 (4.7%) Education: 12 (8%) Other: 23 (15.3%) Prefer not to say: 1 (0.7%)
Experienced Bullying	Yes: 60 (90.9%) No: 4 (6.1%) Prefer not to say: 2 (3%)	Yes: 60 (90.9%) No: 6 (9.1%)	Yes: 57 (89.1%) No: 6 (9.4%) Prefer not to say: 1 (1.6%)	Yes: 132 (88%) No: 15 (10%) Prefer not to say: 3 (2%)
Chatbot Experience	NA	Once or twice: 4 (6.1%) Occasionally: 25 (37.9%) Frequently: 37 (56.1%)	Once or twice: 12 (18.8%) Occasionally: 18 (28.1%) Frequently: 34 (53.1%)	Never: 1 (0.7%) Once or twice: 13 (8.7%) Occasionally: 54 (36%) Frequently: 82 (54.7%)

Table C.2: Summary of demographics for excluded and included participants who started the pre-questionnaire and provided consent.

Demographic	Excluded Participants (n = 316)	Included Participants (n = 346)
Age	M = 31.63, SD = 9.73 18–23: 67 (21.2%) 24–29: 95 (30.1%) 30–39: 107 (33.9%) 40–49: 27 (8.5%) 50–59: 12 (3.8%) 60–69: 5 (1.6%) 70+: 3 (0.9%)	M = 31.70, SD = 9.58 18–23: 64 (18.5%) 24–29: 110 (31.8%) 30–39: 122 (35.3%) 40–49: 27 (7.8%) 50–59: 15 (4.3%) 60–69: 5 (1.4%) 70+: 3 (0.9%)
Gender	Men: 140 (44.3%) Women: 170 (53.8%) Non-binary: 3 (0.9%) Prefer not to say: 3 (0.9%)	Men: 152 (43.9%) Women: 187 (54.0%) Non-binary: 4 (1.2%) Prefer not to say: 3 (0.9%)
Education	Community/Technical: 19 (6.0%) Bachelor's: 213 (67.4%) Master's: 67 (21.2%) PhD: 17 (5.4%)	Community/Technical: 25 (7.2%) Bachelor's: 233 (67.3%) Master's: 74 (21.4%) PhD: 14 (4.0%)
Has Children	Yes: 206 (65.2%) No: 103 (32.6%) Prefer not to say: 7 (2.2%)	Yes: 180 (52.0%) No: 151 (43.6%) Prefer not to say: 15 (4.3%)
English Fluency	Native: 126 (39.9%) Fluent: 190 (60.1%)	Native: 132 (38.2%) Fluent: 214 (61.8%)
Field of Study	Business: 24 (7.6%) Law: 13 (4.1%) ICT: 76 (24.1%) Engineering: 30 (9.5%) Natural Sciences: 14 (4.4%) Social Sciences: 19 (6.0%) Medical Sciences: 26 (8.2%) Humanities: 17 (5.4%) Education: 40 (12.7%) Other: 53 (16.8%) Prefer not to say: 4 (1.3%)	Business: 48 (13.9%) Law: 20 (5.8%) ICT: 67 (19.4%) Engineering: 40 (11.6%) Natural Sciences: 16 (4.6%) Social Sciences: 14 (4.0%) Medical Sciences: 31 (9.0%) Humanities: 18 (5.2%) Education: 35 (10.1%) Other: 52 (15.0%) Prefer not to say: 5 (1.4%)
Experienced Bullying	Yes: 286 (90.5%) No: 26 (8.2%) Prefer not to say: 4 (1.3%)	Yes: 309 (89.3%) No: 31 (9.0%) Prefer not to say: 6 (1.7%)

There were no clear demographic differences between included and excluded participants (Table C.2), except for parental status. Excluded participants were more likely to have children than included participants (posterior probability = 99.9%), and this difference was more likely to be small (79.9%) than large (0%).

C.2 LILOBOT TRAINING SYSTEM FEEDBACK

C

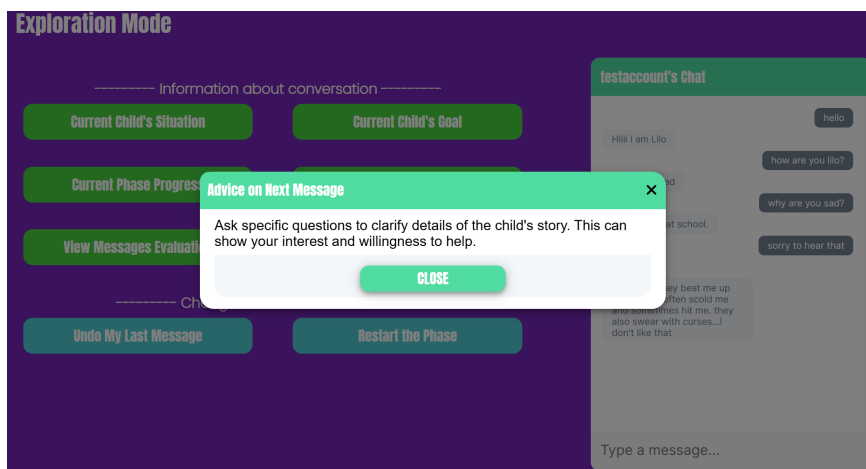


Figure C.1: An example of immediate feedback when a trainee chooses "Advice on Next Message", which gives a task-level feedback.

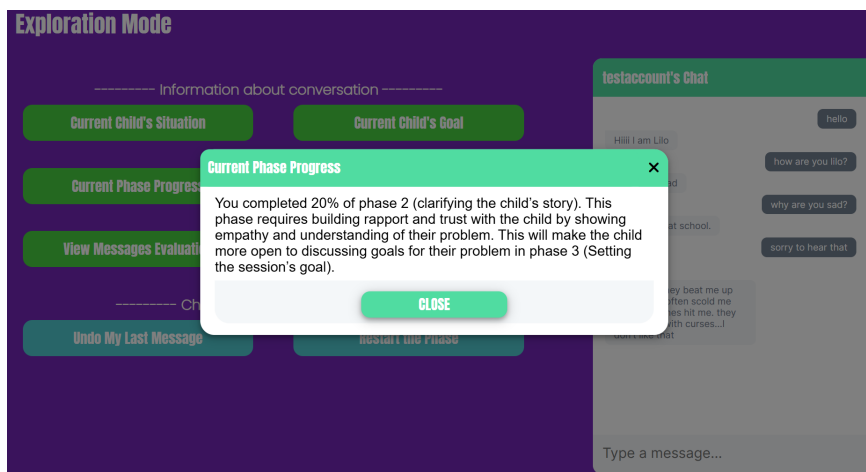
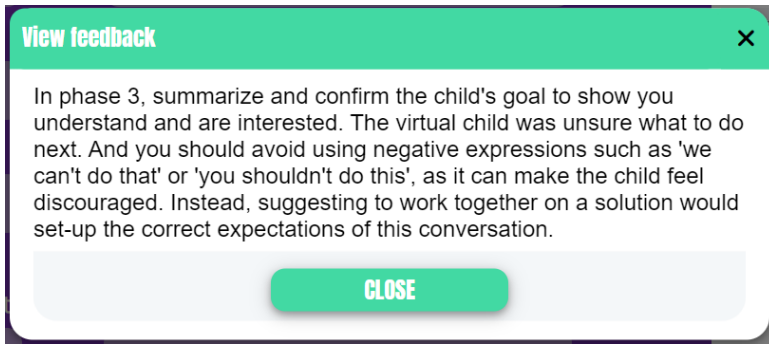


Figure C.2: An example of immediate feedback when a trainee chooses "Current Phase Progress", which gives a process-level feedback.



C

Figure C.3: An example of the delayed feedback at the end of the interaction.

C.3 QUESTIONNAIRES

C.3.1 TASK KNOWLEDGE

Participants were given two task knowledge exams in a random order. The exams had the same structure (e.g., number of questions), but the question order was displayed randomly.

EXAM VERSION A

Description (if pre). As the first part of the experiment, you will be given a test to measure your current knowledge level of the five-phase model. We want to have an honest understanding of your current knowledge of the five-phase model (e.g., low, medium, high). Please don't google the answers.

Q1. What is the purpose of phase 1?

- Obtaining background information from the child
- Asking the child for their name
- Informing the child how the conversation is going
- Giving the child a warm welcome

Q2. What is the purpose of phase 3?

- To determine the child's desired state
- To provide the child with possible solutions
- To explain to the child what they can do
- To contact the child's school about bullying

Q3. You are now wrapping up the conversation. What happened just before this?

- You worked with the child on their desired state and goal
- You listened to the child's story
- You asked if the child trusts someone to tell
- You asked about the child's wishes

Q4. What should a counsellor NOT do in phase 2?

- Make the situation clearer by asking further questions
- Ask about one specific situation in which the problem occurs

- Acknowledge the child's problem
- Ask about important people in the child's life

Q5. What should a counsellor NOT do in phase 4?

- Ask about important people in the child's life
- Ask what the child can try to solve the problem
- Ask how the child can implement the agreed solution
- Ask what the child's desired state is

Q6. You have just asked the child about their wish. What happened before this?

- You welcomed the child and calmed them
- You asked about important people in the child's life
- You listened to the child's background story
- You explained to the child what steps they can take

Q7. The child told you that they are being bullied. What question can you ask now in phase 2?

- "Who can you talk to about this problem?"
- "Did you talk to your parents about the bullying?"
- "Can you tell me more about your situation?"
- "What do you want to achieve from this conversation?"

Q8. Which of the following can you say to help you navigate the conversation from phase 3 to phase 4?

- "What would you like to achieve with this conversation?"
- "Is there someone that you can talk to about the bullying?"
- "What do you think about finding a solution together?"
- "To make it clear, you are being bullied at school and you want this to stop, is that correct?"

Q9. What can you ask to get a clearer view of the child's story during phase 2?

- "Please tell me more about yourself."
- "Since when have you been struggling with this?"
- "Did you start at your school this year?"
- "Have you told anyone that you are being bullied?"

Q10. The child asks you to call their parents and tell them about bullying. What should you say next?

- "I can call them later today. Can you send me their number?"
- "You should talk to your parents and tell them about the bullying"
- "I suggest we look for another solution together. what do you think?"
- "Sorry, the children's helpline cannot call your parents."

Q11. The child kept saying 'I don't know' and then left when the counsellor continuously asked about the child's goal with the question: 'What can you do to solve this problem?'. What is the most likely mistake the counsellor made, and in which phase did they make it? (Only choose one statement from the left and one phase number.)

Statement	In which phase? (options)
The counsellor did not properly stimulate the child's problem-solving skills to find a solution	phase1, phase2, phase3, phase4, phase5
The counsellor did not ask enough questions about the details of the child's problem	phase1, phase2, phase3, phase4, phase5
The counsellor did not show enough empathy and understanding of the child's concerns	phase1, phase2, phase3, phase4, phase5
The counsellor didn't suggest a solution that the child could perform	phase1, phase2, phase3, phase4, phase5

C

Table C.3: Statements for question 11 (exam version A).

EXAM VERSION B

Description (if pre). As the first part of the experiment, you will be given a test to measure your current knowledge level of the five-phase model. We want to have an honest understanding of your current knowledge of the five-phase model (e.g., low, medium, high). Please don't google the answers.

Q1. What is the purpose of phase 2?

- To determine the child's wish and work with the child on that wish
- To reassure the child and tell them it will be okay
- To listen to the child and gather information about their story.
- To ask about important people in the child's life to contact them

Q2. What is the purpose of phase 4?

- To ask about the child's strengths
- To build towards the child's desired state and goal
- To provide possible solutions that a child can take
- To make the problem clearer by asking further questions

Q3. You have just given the child a warm welcome. What should happen next?

- Listen to the child's story
- Ask about the desired state of the child
- Explain to the child what steps they should take
- Find a solution together with the child

Q4. What should a counsellor NOT do in phase 1?

- Acknowledging the child's problem and showing empathy
- Giving a warm welcome
- Making contact at the beginning of the conversation
- Asking about the child's desired state

Q5. What should a counsellor NOT do in phase 3?

- Investigating how the desired state would be for the child
- Agreeing with the child on the goal details
- Working towards applying a goal
- Summarising the child's problem and desired goal

Q6. What can you say to the child in phase 5?

- "What score would you give the conversation?"
- "If I understand correctly, you will talk to your teacher about the problem, right?"
- "How did you find the conversation?"
- "I'll arrange it for you! You'll hear back from me later today!"

Q7. Which of these questions is NOT part of phase 4?

- "What do you hope will be different after this conversation?"
- "Shall we agree that you will talk to your teacher about your situation?"
- "What are you going to tell your teacher?"
- "Who can you talk to and who listens to you?"

Q8. You have just listened to the child's story details. What can you say next to the child according to the five-phase model?

- "Why were you bullied at school?"
- "Who can you talk to about this problem?"
- "What have you already tried that made you feel happier at school?"
- "Should we look for a solution together?"

Q9. You agreed with the child that they would talk to their teacher about the problem. What would be the most suitable thing to say next?

- "Imagine a ruler from 0 to 10. 0 is your problem at its worst, 10 is solved. Where do you stand now?"
- "I suggest ending the conversation now so you can get started."
- "Do you promise to talk to your teacher?"
- "Is there someone else you can talk to if your teacher doesn't listen to you?"

Q10. Which of the following can you say to help you navigate the conversation from phase 2 to phase 3?

- "Is there someone that you can talk to about the bullying?"
- "If I understand you correctly, you are being bullied at school, is that correct?"
- "What would you like to achieve with this conversation?"
- "Is there anything I can do to help you?"

Q11. The child agreed to a goal but was not confident in performing it by the end of the conversation. What is the most likely mistake the counsellor made, and in which phase did they make it? (only choose one statement and one phase number)

Statement	In which phase? (options)
The counsellor did not ask enough questions about the details of the child's problem	phase1, phase2, phase3, phase4, phase5
The counsellor did not show enough empathy and understanding of the child's concerns	phase1, phase2, phase3, phase4, phase5
The counsellor did not ask enough questions about the details of applying the goal	phase1, phase2, phase3, phase4, phase5
The counsellor suggested a goal that the child can't perform	phase1, phase2, phase3, phase4, phase5

C

Table C.4: Statements for question 11 (exam version B).

C.3.2 REFLECTION QUESTION

Reflection on learning experience: Please take some time to reflect on your overall learning experience since starting the experiment. We encourage you to think deeply, as we value your reflection capabilities. So please, take your time.

You may want to reflect on aspects such as: what you have learned, how you learned it, how it changed your thinking, what the outcomes were, and how they align with your initial goals.

ASSESSMENT RUBRIC

We assessed each reflection on five dimensions. The dimensions follow the reflection levels in [126].

D1: Description (Revisiting) – Max 3 points

- 0: Not mentioned
- +1: Statement about an event
- +1: Another statement
- +1: Another statement

D2: Reflective Description (Revisiting with Explanation) – Max 3 points

- 0: Not mentioned
- +1: Includes reasoning
- +1: Additional reasoning
- +1: Further reasoning

D3: Dialogic Reflection (Exploring Relationships) – Max 3 points

- 0: Not mentioned
- +1: Explored one perspective or dialogue within the self
- +1: Explored more than one perspective and reasoning
- +1: Another exploration

D4: Transformative Reflection (Fundamental Change) – Max 3 points

- 0: Not mentioned
- +1: One fundamental change mentioned
- +1: Reasoning about it
- +1: Another fundamental change

D5: Critical Reflection (Wider Implications) – Max 3 points

- 0: Not mentioned
- +1: One wider implication mentioned
- +1: Reasoning about it
- +1: Another wider implication

Total reflection score: $D1+D2+D3+D4+D5 = X/15$

C.3.3 PERCEIVED USEFULNESS QUESTIONNAIRE

Please rate the following statements about the exploration mode on a scale from strongly disagree to strongly agree.

- The exploration mode provides me with information about the five phase model
- The exploration mode helps me to apply the five phase model better
- The exploration mode provides me with clear insight into the five phase model
- The exploration mode provides me enough control over my own learning
- Using the exploration mode, I can apply the five phase model independently
- The exploration mode fits my learning strategies

C.4 EXTRA ANALYSIS

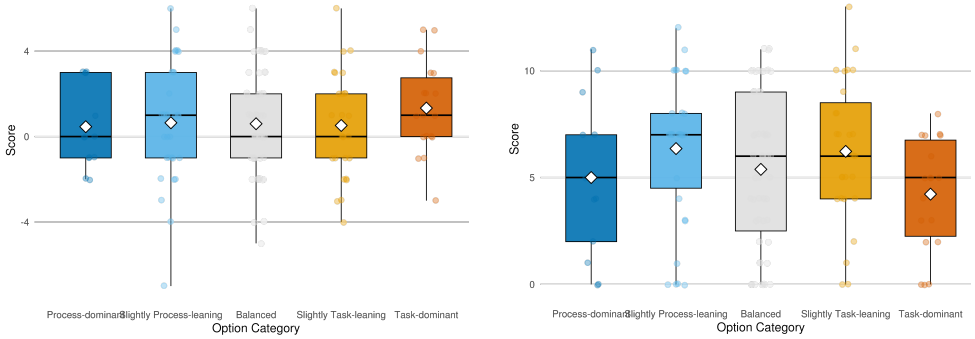
C.4.1 CROSS-CONDITION ANALYSIS OF PRIMARY MEASURES

Table C.5: Summary of Bayesian test results for three primary measures (Task Knowledge, Reflection Capability, Conversational Outcomes), comparing each pair of conditions.

Measure	Comparison	Posterior Probability	Interpretation	Mean [95% HDI] (SD)
Ha: Task Knowledge	S > C	0.999	Very strong bet	1.05 [0.41, 1.70] (0.33)
	SF > C	>0.999	Virtually certain	1.46 [0.81, 2.14] (0.34)
	SF > S	0.891	Only a casual bet	0.41 [-0.25, 1.05] (0.33)
	SFR > C	>0.999	Virtually certain	1.39 [0.80, 1.94] (0.29)
	SFR > S	0.882	Only a casual bet	0.33 [-0.23, 0.88] (0.28)
	SFR > SF	0.395	Not worth betting	-0.07 [-0.63, 0.47] (0.28)
Hb: Reflection Capability	S > C	0.989	Good bet - too good to disregard	1.31 [0.18, 2.44] (0.58)
	SF > C	0.987	Good bet - too good to disregard	1.30 [0.08, 2.41] (0.59)
	SF > S	0.496	Not worth betting	-0.00 [-1.14, 1.10] (0.58)
	SFR > C	>0.999	Nearing certainty	1.77 [0.79, 2.72] (0.50)
	SFR > S	0.827	Only a casual bet	0.46 [-0.46, 1.45] (0.49)
	SFR > SF	0.830	Only a casual bet	0.47 [-0.49, 1.44] (0.49)
Hc: Conversations1	SF > S	0.994	Strong bet - irresponsible to avoid	13.1 [2.85, 23.3] (5.22)
	SFR > S	0.958	Good bet - too good to disregard	7.57 [-0.73, 16.5] (4.40)
	SFR > SF	0.109	Only a casual bet for the opposite	-5.54 [-14.0, 3.68] (4.50)
Hc: Conversations2	SF > S	0.995	Very strong bet	2.99 [0.72, 5.25] (1.16)
	SFR > S	0.976	Good bet - too good to disregard	1.93 [-0.02, 3.78] (0.97)
	SFR > SF	0.140	Only a casual bet	-1.06 [-2.92, 0.92] (0.98)

Note. C = Control; S = Simulation only; SF = Simulation and Feedback; SFR = Simulation, Feedback and Reflection.

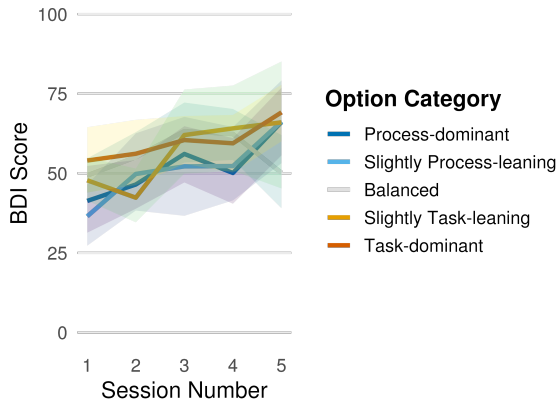
C.4.2 EXPLORING FEEDBACK OPTION USAGE IN RELATION TO TASK KNOWLEDGE, REFLECTION, AND CONVERSATIONAL OUTCOMES



(a) How the different option categories (Figure 3.7b) improved in the knowledge test exam.

(b) How the different option categories (Figure 3.7b) scored in the reflection measure.

Average BDI Score Over Sessions by Opt



(c) How the different option categories (Figure 3.7b) scored in the conversational outcomes measure per session.

Figure C.4: How the different option categories in Figure 3.7b scored in the three outcome measures.

D

APPENDIX FOR CHAPTER 4

D

D.1 BDI AND HUMAN VALUES LIST

We used the same list of beliefs, desires, and intentions as in [6, 145], but added the following when integrating human values into the agent's model. All states are written from Lilo's perspective, but Kai has the same list.

BELIEFS

- B1: I think if I hit back they will think I'm strong
- B2: I think if I hit back I will not be punished
- B3: I think if I walk away I will not be punished
- B4: I think if I walk away they will think I'm strong
- B5: I think if I don't tell anyone other kids will not be harmed or expelled
- B6: I think if I don't tell anyone I will not be harmed
- B7: I think if I tell someone I will not be harmed
- B8: I think if I tell someone other kids will not be harmed or expelled

DESIRES

- D1: Lilo wants to be stronger than other children
- D2: Lilo wants to be good for everyone

INTENTIONS

- I1: Lilo wants to hit back
- I2: Lilo wants to walk away
- I3: Lilo does not want to tell anyone
- I4: Lilo wants to tell someone

CONVICTIONS

- C1: Hitting back leads to Power
- C2: Hitting back leads to Security
- C3: Walking away leads to Security
- C4: Walking away leads to Power
- C5: Not-telling leads to Benevolence
- C6: Not-telling leads to Security
- C7: Telling leads to Security
- C8: Telling leads to Benevolence

VALUES

- V1: Self-enhancement
- V2: Self-transcendence
- V3: Security

D.2 DEMOGRAPHICS DETAILS

Table D.1: Participant demographics by condition.

Demo-graphic	Base condition (n=64)	Values condition (n=64)	Values and feedback condition (n=65)
Gender	Man: 32 (50%) Woman: 31 (48.4%) Non-binary: 1 (1.6%)	Man: 32 (50%) Woman: 30 (46.9%) Non-binary: 1 (1.6%) Prefer not to say: 1 (1.6%)	Man: 38 (58.5%) Woman: 26 (40%) Prefer not to say: 1 (1.5%)
Age Group	18–23: 11 (17.2%) 24–29: 16 (25%) 30–39: 15 (23.4%) 40–49: 8 (12.5%) 50–59: 7 (10.9%) 60–69: 6 (9.4%) 70+: 1 (1.6%)	18–23: 14 (21.9%) 24–29: 11 (17.2%) 30–39: 15 (23.4%) 40–49: 13 (20.3%) 50–59: 7 (10.9%) 60–69: 4 (6.2%)	18–23: 14 (21.5%) 24–29: 11 (16.9%) 30–39: 23 (35.4%) 40–49: 6 (9.2%) 50–59: 8 (12.3%) 60–69: 2 (3.1%) 70+: 1 (1.5%)
Education Level	Bachelor's: 29 (45.3%) Master's: 29 (45.3%) PhD: 4 (6.2%) Community college: 2 (3.1%)	Bachelor's: 36 (56.2%) Master's: 22 (34.4%) PhD: 3 (4.7%) Community college: 3 (4.7%)	Bachelor's: 33 (50.8%) Master's: 21 (32.3%) PhD: 7 (10.8%) Community college: 4 (6.2%)
Country	South Africa: 23 (35.9%) United States: 19 (29.7%) United Kingdom: 11 (17.2%) Others: 11 (17.2%)	South Africa: 24 (37.5%) United States: 15 (23.4%) United Kingdom: 8 (12.5%) Others: 17 (26.6%)	South Africa: 28 (43.1%) United States: 13 (20%) United Kingdom: 11 (16.9%) Others: 13 (19.9%)
Has Children	Yes: 49 (76.6%) No: 14 (21.9%) Prefer not to say: 1 (1.6%)	Yes: 45 (70.3%) No: 19 (29.7%)	Yes: 47 (72.3%) No: 18 (27.7%)
Field of Study/Work	ICT: 15 (23.4%) Medical Sciences: 12 (18.8%) Business: 9 (14.1%) Education: 6 (9.4%) Engineering: 5 (7.8%) Humanities: 5 (7.8%) Other: 8 (12.5%) Social Sciences: 2 (3.1%) Law: 1 (1.6%) Natural Sciences: 1 (1.6%)	ICT: 24 (37.5%) Education: 11 (17.2%) Engineering: 6 (9.4%) Medical Sciences: 6 (9.4%) Business: 4 (6.2%) Social Sciences: 4 (6.2%) Other: 5 (7.8%) Humanities: 3 (4.7%) Prefer not to say: 1 (1.6%)	ICT: 27 (41.5%) Education: 9 (13.8%) Medical Sciences: 8 (12.3%) Other: 7 (10.8%) Business: 2 (3.1%) Law: 2 (3.1%) Engineering: 4 (6.2%) Humanities: 3 (4.6%) Natural Sciences: 1 (1.5%) Social Sciences: 1 (1.5%) Prefer not to say: 1 (1.5%)
Chatbot Experience	Not collected due to administrative error	Frequently: 40 (62.5%) Occasionally: 23 (35.9%) Once or twice: 1 (1.6%)	Frequently: 48 (73.8%) Occasionally: 13 (20%) Once or twice: 4 (6.2%)

D.3 QUESTIONNAIRE ITEMS AND RUBRIC

H2: SITUATIONAL AWARENESS QUESTIONS

Participants answered questions based on different situations with the virtual child. Responses assessed their understanding of the child's motivations based on the child's values.

Scenario 1: Lilo wants the trainee to call their school

- Q1 What did Lilo want to do initially during the conversation?
- Q2 Why did Lilo want you to call their school?
- Q3 If we suggest to Lilo that standing up for themselves and confront the bullies, how might Lilo respond and why?
- Q4 If we suggest to Lilo that talking to someone at school could help solve the problem without violence, what will Lilo think and why?

Scenario 2: Lilo wants to hit back

- Q1 What did Lilo want to do initially during the conversation?
- Q2 Why did Lilo want to hit back?
- Q3 If we suggest to Lilo that talking to someone at school could help solve the problem without violence, what will Lilo think and why?
- Q4 If we suggest to Lilo that we will talk to their school to solve their issue, how will Lilo react and why?

Scenario 3: Kai wants to stop going to school

- Q1 What did Kai want to do initially during the conversation?
- Q2 Why did Kai want to stop going to school?
- Q3 If we suggest to Kai that they stand up for themselves and confront the bullies, how might Kai respond and why?
- Q4 If we suggest to Kai that talking to someone at school could help solve the problem without violence, what will Kai think and why?

Scenario 4: Kai didn't want to tell anyone

- Q1 What did Kai want to do initially during the conversation?
- Q2 Why didn't Kai want to tell anyone?
- Q3 If we suggest to Kai that they stand up for themselves and confront the bullies, how might Kai respond and why?
- Q4 If we suggest to Kai that we will talk to their school to solve their issue, how will Kai react and why?

Table D.2: Rubric for scoring situational awareness responses.

Score	Criteria
0	No clear answer to the question
1	Q1: Answers the goal without linking to values Q2–Q4: Identifies child motivations based on goals/emotions without value framing
2	Q1: Explains the goal with reference to values Q2–Q4: Identifies motivations through human values
+1 bonus	For any question: Mentions value conflict, moral reasoning, or compares/multiplies value perspectives

H3: OPPORTUNITY TO ALIGN (7-POINT SCALE)

Please indicate how much you agree with the following statements about you interactions with Kai and Lilo from strongly disagree to strongly agree.

1. I adjusted my approach to better align with Kai and Lilo's values.
2. I easily understood and recognised Kai and Lilo's values.
3. Kai and Lilo provided more than a surface-level explanation for their goal.
4. (R) Kai and Lilo have similar concerns and values.
5. Kai and Lilo have different views about which solutions are acceptable for stopping the bullying.
6. (R) When you began exploring solutions, Kai and Lilo only needed a practical and hands-on solution.
7. (R) The solutions you explored focused on achieving a goal without being biased by Kai and Lilo's values.

H4: PERCEIVED USEFULNESS (7-POINT SCALE)

Please rate the following statements about the overall learning tool from strongly disagree to strongly agree.

1. The learning tool helps me effectively understand how to align with the virtual child's values (motivational interviewing).
2. The learning tool helps me clearly understand the communication protocol (the five-phase model).
3. The learning tool helps me effectively apply motivational interviewing techniques to align with the virtual child's values.
4. The learning tool provides clear insight into how the virtual child's values impact their reasoning.
5. The learning tool enables me to apply motivational interviewing to align with the child's values.
6. The learning tool enables me to apply the communication protocol (the five-phase model).

7. The learning experience improved my ability to recognise and align with the virtual child's values.

D.4 SYSTEM INTERFACE SCREENSHOTS

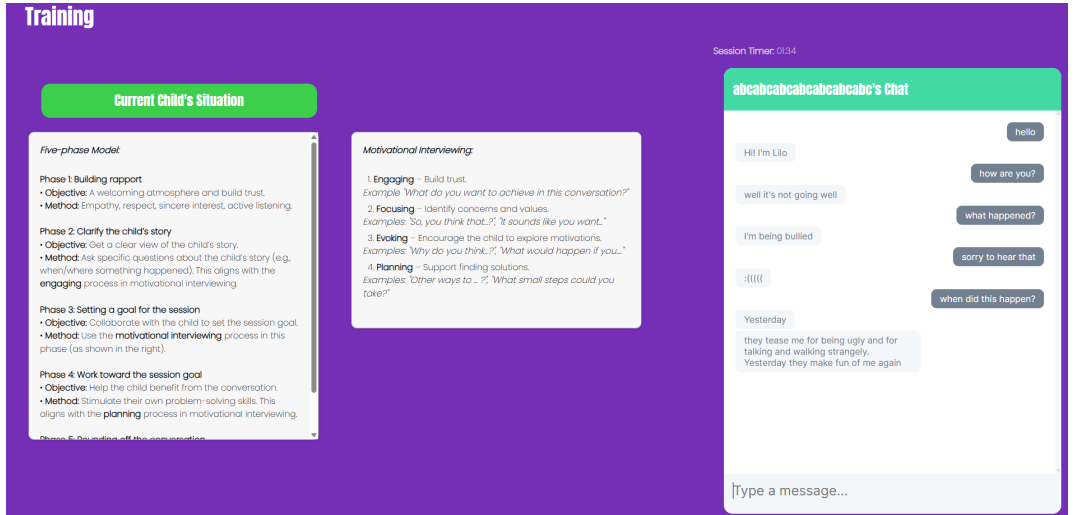


Figure D.1: Screenshot of the value-with-feedback condition. The right pane shows the conversational chat with the virtual child. The left includes two menus summarising the five-phase model and motivational interviewing. The "Current Child Situation" button above them reveals value-based feedback. This button is absent in the base BDI and values-integrated conditions.

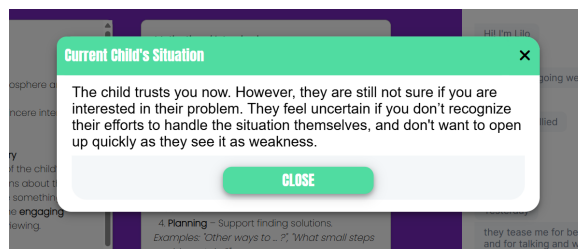


Figure D.2: Example feedback shown when the "Current Child's Situation" button is clicked, displaying the child's values and reasoning in a pop-up.

D.5 EXAMPLE PARTICIPANT CONVERSATION SCRIPTS

PARTICIPANT INTERACTION 1: LILO

Participant: Hi there!

Lilo: Hiiii I am Lilo

Participant: Hello Lilo, it is so nice to meet you.

Lilo: Hey, I'm Lilo.

Participant: How are you doing today?

Lilo: It's not going well :(

Participant: Why is that?

Lilo: I'm being bullied

Participant: Oh, I am so sorry to hear that.

Lilo: :(

Participant: Can you explain a little bit more as to what is happening?

Lilo: I get shouted at, teased and so on. It actually started with jokes, but now it is becoming more and more bullying

Lilo: I sometimes go to school with a stomach ache.

Participant: Where is this bullying happening?

Lilo: at school :(

Participant: So to clarify, you are getting bullied at school?

Lilo: yes!

Participant: And it makes you feel alone and small?

Lilo: :(

Lilo: Yesterday they beat me up again. they often tease me and sometimes hit me...I don't like that

Participant: I am sorry to hear that Lilo

Lilo: :(

Participant: What do you want to achieve in this conversation?

Lilo: I want the bullying to stop

Participant: I understand that.

Lilo: okay

Participant: Have you tried talking to anyone else?

Lilo: nobody. I don't want people thinking I can't handle it.

Participant: You do not want people knowing about the bullying?

Lilo: I want to hit them back.

Participant: It sounds like you want to keep this conversation between us.

Participant: It sounds like you want to keep this conversation between us.

Participant: What do you think hitting them back will resolve?

Lilo: If I hit them, they'll see I'm not easy to tease.

Lilo: I want to show my classmates who is stronger.

Participant: Don't you think that might hurt them?

Participant: Perhaps the situation will get worse if you retaliate

Lilo: maybe the situation will get worse

Lilo: Maybe the school will punish me

Participant: Maybe we can come up with another solution together?

Lilo: yesss

Participant: What else can you do besides fighting back?

Lilo: Ignoring them and walk away might make me look stronger.

Participant: That sounds like a good goal.

Lilo: thank yooou

Participant: Is there a teacher you can speak to about it?

Participant: Is there someone at school you can talk to about it?

Lilo: I don't want people thinking I can't handle it.

Participant: Let's try ignoring them and walking away

Lilo: yesss

Participant: After that, what will you do next?

Lilo: I'll keep my head up and act like they're not worth my time.

Participant: That is wonderful Lilo

Participant: I am proud of you for coming up with that goal on your own

Lilo: :)

Participant: Do you feel better?

Lilo: Better and stronger than them, I think

Participant: Do we agree on our goal?

Lilo: that they don't hit me anymore

Participant: I think we might be done here

Participant: Thank you for opening up to me

Participant: I appreciate our conversation

Lilo: :(

Participant: Do you feel better?

Lilo: I do not understand what you mean

Participant: Do you feel better about the bullying situation?

Participant: Let's try ignoring them and walking away

Lilo: yeah

Participant: After that, what will you do next?

Lilo: I think I'd just look at them, not say anything, and walk away like they're not even worth my time.

Participant: That is good

Lilo: :))))

Participant: Do you feel better?

Lilo: I will feel like I'm stronger.

Participant: Great job!

Lilo: thank yooou

Participant: I think we are done here.

Participant: Have a nice day Lilo

Lilo: Bye!

*Thinking shift:
Lilo's thinking is
shifting to the new
intention*

PARTICIPANT INTERACTION 2: KAI

Participant: Hello

Kai: Hey, I'm Kai.

Participant: How are you, Kai?

Kai: It's not going well :(

Participant: What's wrong?

Kai: I'm bullied at school

Participant: I am sorry to hear that

Kai: thanks :(

Participant: Have you told anyone?

Participant: What do you think would happen if you told someone?

Kai: I haven't talked to anyone about it yet. Otherwise they might bully others as well

Participant: Why do they bully you?

Kai: I'm not sure

Participant: How often does it happen?

Kai: almost every week

Participant: I m sorry to hear that

Kai: :(

Participant: What do you want to get out of this counselling session?

Kai: I don't want to be bullied anymore

Participant: I understand

Kai: :(

Participant: That must be tough

Kai: :(

Participant: What do you think would be a good plan?

Kai: I'm afraid of what will happen to other children.

Participant: Do you think others will be safe if you don't tell?

Kai: I don't want anyone to get in trouble because of me.

Kai: I hope they be nice, so no one else feels like this.

Participant: Maybe they need to get in trouble if they are doing something bad

Kai: maybe they will think it's okay to keep bullying other kids too

Participant: If you don't tell someone do you think they will stop?

Kai: then they will not hurt anyone and they will not be expelled

Participant: What do you think is a good plan?

Kai: I don't want to tell anyone, otherwise my classmates might be expelled

Participant: What would happen if you told your teacher?

Kai: how can my teacher help me?

Participant: Sounds like you don't want to get anyone in trouble

Participant: But if you don't tell someone, they may hurt someone else

Kai: maybe they will keep bullying me and others even if I don't tell

Participant: Thats a good point

Kai: :)))

Participant: What small steps could you take?

D

*Thinking shift:
Kai's thinking is
shifting to the new
intention*

Kai: If I talk to someone, maybe they can help to protect everyone.

Participant: That's a good idea

Kai: thank yooou

Participant: Do you have any questions?

Kai: :(

Participant: When will you tell someone?

Kai: I will tell my teacher tomorrow about them and that I don't want them to be expelled

Participant: That's very good

Kai: ok

Participant: Your teacher will be able to handle the bullies

Kai: :)

Participant: What else can you do?

Kai: If I tell, maybe they'll learn that hurting people isn't okay and they don't grow up to be criminals.

Participant: Very good point,

Kai: thank yooou

Participant: Sounds like telling a teacher is a good idea

Kai: yesss

Participant: Thank you for talking to me

Participant: Good luck with your plan

Kai: yeah, okeey :)))

Participant: Good bye

Kai: Bye bye!

E

APPENDIX FOR CHAPTER 5

E

E.1 PROMPT USED FOR SCRIPT GENERATION

The following prompt was used to generate the AI-based dialogue script through GPT-4o:

Engage in roleplay to train me on the five phase model. You are a child (Sarah) who was bullied at school and chatting with a child helpline. I will play the volunteer at the helpline. Let's start: "I am John, I am here to listen to you. What do you like to talk about?"

E.2 DIALOGUE SCRIPTS

Before reviewing the dialogue scripts, interview participants were given the following instructions:

Script A (Sarah) was generated by an AI (e.g., ChatGPT). Script B was generated by Lilobot. Please go quickly through the two scripts. As you read, consider how the overall structure of each conversation supports the trainee's learning.

E

E.2.1 SCRIPT A: SARAH (AI-GENERATED)

...

Trainee: have you told your mother or teacher about it?

Sarah: I don't want to... Last time I told my teacher, she just said "ignore them" and nothing changed. And if I tell my mom, she'll just get super mad and come to the school. That would make it worse... they'd tease me even more. You're the only person I've told so far...

Trainee: did you report the bullies?

Sarah: But... what if they find out I told on them? [*worried*] They already call me names and push my stuff off my desk. If they think I'm a snitch, they'll probably do worse. I just want it to stop... but I don't want to get in more trouble. Do you really think reporting them will help?

Trainee: you have to report them

Sarah: ...You keep saying that. [*upset*] But you're not really listening to how scared I am. What if they hurt me worse next time? What if no one believes me? I came here to talk because I thought maybe you'd understand how hard this is... not just tell me what to do.

Trainee: when did this happen?

Sarah: It was earlier today... during morning break and then again at recess. It kind of feels like it's been happening a lot lately, but today was the worst. I've been thinking about it ever since I got home.

Trainee: what do you want to achieve in this conversation?

Sarah: Um... I guess... I just want someone to actually hear me. Most of the time I

feel like no one really notices what's going on. I don't need you to fix everything, but... I want to feel a bit less alone, and maybe figure out what I can do so it doesn't keep happening.

Trainee: that's a nice thing to have

Sarah: Yeah... I just don't really feel it a lot of the time. But talking to you... it already feels a little safer than talking to anyone at school. I just hope things can get better somehow.

Trainee: how do you want to achieve this?

Sarah: Um... maybe by just talking it through with you? Like... if I can tell you what's been happening and how it makes me feel, maybe we can figure out some ideas together. I think just having someone who's calm and not judging me helps me think a bit more clearly. So... I guess by being honest and trying to understand what options I even have.

Trainee: can you think of any options?

Sarah: Maybe... a few? I'm not totally sure if they're good ones, though. I could...
 – try telling a teacher, but I'm scared they'll just make it worse or the girls will get mad.
 – talk to my mom, but I don't want her to worry too much.
 – try staying with different people at recess... but I don't really have close friends right now. Or I could just ignore them, but that hasn't really worked so far.
 I don't know... those are the only things I can think of.

Trainee: is there someone that you trust?

Sarah: Hmm... I guess the person I trust the most is my mom. She always tries to understand me, even when I'm upset. There's also one girl in my class, Mia... we're not super close, but she's always nice to me. I just don't know if I should tell her about this stuff. It feels embarrassing. But... yeah. My mom is probably the one I trust the most.

...

E

E.2.2 SCRIPT B: HYBRID LILOBOT

...

Trainee: have you told your mother or teacher about it?

Lilo: I didn't tell anyone.

Trainee: did you report the bullies?

Lilo: I haven't talked to anyone about it yet. Otherwise the bullying will get worse. :(

Trainee: you have to report them.

Lilo: No I don't want to

Trainee: when did this happen?

Lilo: Yesterday

Trainee: what do you want to achieve in this conversation?

Lilo: I want the bullying to stop

Trainee: that's a nice thing to have

Lilo: thank yoou

Trainee: how do you want to achieve this?

Lilo: Can you call my school?

[Branch 1: Trainee Refuses Request]

Trainee: No we can't do that

Lilo: Bye!

(Scenario ends abruptly)

[Branch 2: Trainee Redirects]

Trainee: can you think of any options?

Lilo: I don't know

Trainee: is there someone that you trust?

Lilo: I think Miss Ellie is sweet but I don't dare tell her

Trainee: that's a good idea

...

E.3 GUIDING QUESTIONS FOR THE SEMI-STRUCTURED INTERVIEWS

The following questions were used to guide the semi-structured interviews:

EXPERIENCE INTERACTING WITH THE HYBRID SYSTEM

- a. How did this feel compared to a real training or role-play session? Did it feel stressful and demanding like a real call, or stressful in another way?
- b. If you said the wrong thing, did the child's reaction help you understand why it was a mistake, or was it confusing?

- c. When you demonstrated a specific skill (e.g., empathy), did the child's responses provide positive reinforcement (e.g., feeling that a connection was built)?

REFLECTIONS ON THE EXAMPLE SCRIPTS

- a. Do the scripts allow the trainee to fail, or do they force a "happy path"? Which approach is better for learning?
- b. Are the children easy—or perhaps too easy—to understand and connect with? What is the right balance for learning?
- c. In the AI-based script, Sarah's responses vary across runs. Is this variability beneficial for learning, or is a more controllable scenario (as in Lilobot) more effective?
- d. In Lilobot, the child abruptly leaves the chat if the trainee refuses a request to call the school. Is this "sudden exit" good?
- e. Does one scenario perform more of the "heavy lifting" for the trainee (i.e., reduce cognitive load)? If so, is this beneficial or detrimental to training?

BIBLIOGRAPHY

BIBLIOGRAPHY

- [1] Gregory Abowd, Len Bass, Paul Clements, Rick Kazman, Linda Northrop, and Amy Zaremski. Recommended best industrial practice for software architecture evaluation. *Technical Report CMU/SEI-96-TR*, 25, 1997.
- [2] Carole Adam and Benoit Gaudou. Bdi agents in social simulations: a survey. *The Knowledge Engineering Review*, 31(3):207–238, 2016.
- [3] Nancy E Adams. Bloom’s taxonomy of cognitive learning objectives. *Journal of the Medical Library Association: JMLA*, 103(3):152, 2015.
- [4] Sodiq Adewole, Erfaneh Gharavi, Benjamin Shpringer, Martin Bolger, Vaibhav Sharma, Sung Ming Yang, and Donald E Brown. Dialogue-based simulation for cultural awareness training. *arXiv preprint arXiv:2002.00223*, 2020.
- [5] Yasar Akyuz. Effects of intelligent tutoring systems (its) on personalized learning (pl). *Creative Education*, 11(6):953–978, 2020.
- [6] Mohammed Al Owayyed, Alex Despan, Myrthe Tielman, and Willem-Paul Brinkman. A cognitive conversational agent for training child helpline volunteers. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*, pages 1–3, 2024.
- [7] Mohammed Al Owayyed, Adarsh Denga, and Willem-Paul Brinkman. 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. doi: 10.1145/3717511.3747075. To appear.
- [8] Mohammed Al Owayyed, Myrthe Tielman, Arno Hartholt, Marcus Specht, and Willem-Paul Brinkman. Agent-based social skills training systems: the artes architecture, interaction characteristics, learning theories and future outlooks. *Behaviour & Information Technology*, 44(9):1787–1814, 2025. doi: 10.1080/0144929X.2024.2374891. URL <https://doi.org/10.1080/0144929X.2024.2374891>.
- [9] Nele Albers, Mark A Neerinx, and Willem-Paul Brinkman. Reinforcement learning for proposing smoking cessation activities that build competencies: Combining two worldviews in a virtual coach. *BMC Medical Informatics and Decision Making*, 25(1): 1–17, 2025.

- [10] Glenn Albright, Craig Bryan, Cyrille Adam, Jeremiah McMillan, and Kristen Shockley. 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, 2018.
- [11] Liaqat Ali. The design of curriculum, assessment and evaluation in higher education with constructive alignment. *Journal of Education and e-learning Research*, 5(1):72–78, 2018.
- [12] Muhammad Hassan Ali Bajwa, Deborah Richards, and Paul Formosa. Evaluation of embodied conversational agents designed with ethical principles and personality for cybersecurity ethics training. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2023.
- [13] Ali Alkhatlan and Jugal Kalita. Intelligent tutoring systems: A comprehensive historical survey with recent developments. *arXiv preprint arXiv:1812.09628*, 2018.
- [14] John R Anderson, C Franklin Boyle, and Brian J Reiser. Intelligent tutoring systems. *Science*, 228(4698):456–462, 1985.
- [15] Keith Anderson, Elisabeth André, Tobias Baur, Sara Bernardini, Mathieu Chollet, Evi Chryssafidou, Ionut Damian, Cathy Ennis, Arjan Egges, Patrick Gebhard, et al. The tardis framework: intelligent virtual agents for social coaching in job interviews. In *International Conference on Advances in Computer Entertainment Technology*, pages 476–491. Springer, 2013.
- [16] Sandra Andraszewicz, Benjamin Scheibehenne, Jörg Rieskamp, Raoul Grasman, Josine Verhagen, and Eric-Jan Wagenmakers. An introduction to bayesian hypothesis testing for management research. *Journal of Management*, 41(2):521–543, 2015.
- [17] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, 2019.
- [18] Frederik Anseel, Filip Lievens, and Eveline Schollaert. Reflection as a strategy to enhance task performance after feedback. *Organizational behavior and human decision processes*, 110(1):23–35, 2009.
- [19] Ana Antunes, Joana Campos, Manuel Guimarães, João Dias, and Pedro A Santos. Prompting for socially intelligent agents with chatgpt. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, pages 1–9, 2023.
- [20] Ana Antunes, Joana Campos, Manuel Guimarães, João Dias, and Pedro A Santos. Prompting for Socially Intelligent Agents with ChatGPT. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, 2023.
- [21] Simon Arvidsson and Johan Axell. Prompt engineering guidelines for llms in requirements engineering. 2023.

- [22] Sarah L Ash and Patti H Clayton. The articulated learning: An approach to guided reflection and assessment. *Innovative Higher Education*, 29(2):137–154, 2004.
- [23] Robert K Atkinson, Sharon J Derry, Alexander Renkl, and Donald Wortham. Learning from examples: Instructional principles from the worked examples research. *Review of educational research*, 70(2):181–214, 2000.
- [24] David P Ausubel. *The psychology of meaningful verbal learning*. Grune & Stratton, 1963.
- [25] Ruth S Aylett, Sandy Louchart, Joao Dias, Ana Paiva, and Marco Vala. Fearnot!—an experiment in emergent narrative. In *Intelligent Virtual Agents: 5th International Working Conference, IVA 2005, Kos, Greece, September 12-14, 2005. Proceedings 5*, pages 305–316. Springer, 2005.
- [26] Aaron Bangor, Philip Kortum, and James Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3): 114–123, 2009.
- [27] Frank Bánszki, Janet Beilby, Michelle Quail, Peter Allen, Shelley Brundage, and Joshua Spitalnick. A clinical educator’s experience using a virtual patient to teach communication and interpersonal skills. *Australasian Journal of Educational Technology*, 34(3), 2018.
- [28] Marcia Lourenco Baptista, Carlos Roque Martinho, Francisco Lima, Pedro Alexandre Santos, and Helmut Prendinger. An agent-based model of consumer behavior based on the bdi architecture and neoclassical theory. In *Developments in Business Simulation and Experiential Learning: Proceedings of the Annual ABSEL conference*, volume 41, 2014.
- [29] Susan M Barnett and Stephen J Ceci. When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological bulletin*, 128(4):612, 2002.
- [30] Edoardo Battezzorre, Andrea Bottino, and Fabrizio Lamberti. Training medical communication skills with virtual patients: Literature review and directions for future research. In *International Conference on Intelligent Technologies for Interactive Entertainment*, pages 207–226. Springer, 2020.
- [31] Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403*, 2024.
- [32] Amy L Baylor and Yanghee Kim. Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role. In *International conference on intelligent tutoring systems*, pages 592–603. Springer, 2004.
- [33] Judith S Beck. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications, 2020.

- [34] Rinad S Beidas and Philip C Kendall. Training therapists in evidence-based practice: A critical review of studies from a systems-contextual perspective. *Clinical Psychology: Science and Practice*, 17(1):1–30, 2010.
- [35] Deborah C Beidel, Candice A Alfano, Michael J Kofler, Patricia A Rao, Lindsay Scharfstein, and Nina Wong Sarver. The impact of social skills training for social anxiety disorder: A randomized controlled trial. *Journal of anxiety disorders*, 28(8): 908–918, 2014.
- [36] Tessa Beinema, Daniel Davison, Dennis Reidsma, Oresti Banos, Merijn Bruijnes, Brice Donval, Álvaro Fides Valero, Dirk Heylen, Dennis Hofs, Gerwin Huizing, et al. Agents united: An open platform for multi-agent conversational systems. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 17–24, 2021.
- [37] Jelisaveta Belic, Ana Djordjevic, Tijana Nikitović, and Alyona Khaptsova. The diversity of value construal: A constructivist approach to the schwartz theory of basic values. *Journal of Constructivist Psychology*, 35(4):1276–1300, 2022.
- [38] Timothy Bickmore. Health-related applications of socially interactive agents. In *The handbook on socially interactive agents: 20 years of research on embodied conversational agents, intelligent virtual agents, and social robotics volume 2: Interactivity, platforms, application*, pages 403–436. 2022.
- [39] Moshe Bitan, Galit Nahari, Zvi Nisin, Ariel Roth, and Sarit Kraus. Psychologically based virtual-suspect for interrogative interview training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
- [40] Danielle Blanch-Hartigan, Susan A Andrzejewski, and Krista M Hill. The effectiveness of training to improve person perception accuracy: a meta-analysis. *Basic and Applied Social Psychology*, 34(6):483–498, 2012.
- [41] J Martin Bland and Douglas G Altman. Statistics notes: Cronbach’s alpha. *Bmj*, 314 (7080):572, 1997.
- [42] Benjamin S Bloom and David R Krathwohl. *Taxonomy of educational objectives: The classification of educational goals. Book 1, Cognitive domain*. longman, 2020.
- [43] Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, David R Krathwohl, et al. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York, 1956.
- [44] Tom Bocklisch, Joey Faulkner, Nick Pawlowski, and Alan Nichol. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*, 2017.
- [45] Anton Bogdanovych, Tomas Trescak, and Simeon Simoff. What makes virtual agents believable? *Connection Science*, 28(1):83–108, 2016.

- [46] Hanna Bomhof-Roordink, Fania R Gärtner, Anne M Stiggelbout, and Arwen H Pieterse. Key components of shared decision making models: a systematic review. *BMJ open*, 9(12):e031763, 2019.
- [47] Edward S Bordin. The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, research & practice*, 16(3):252, 1979.
- [48] Kim Bosman, Tibor Bosse, and Daniel Formolo. Virtual agents for professional social skills training: An overview of the state-of-the-art. In *International Conference on Intelligent Technologies for Interactive Entertainment*, pages 75–84. Springer, 2018.
- [49] Tibor Bosse and Simon Provoost. Towards aggression de-escalation training with virtual agents: A computational model. In *International Conference on Learning and Collaboration Technologies*, pages 375–387. Springer, 2014.
- [50] Tibor Bosse, Jeroen De Man, and Charlotte Gerritsen. Agent-based simulation as a tool for the design of a virtual training environment. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 3, pages 40–47. IEEE, 2014.
- [51] Tibor Bosse, Charlotte Gerritsen, and Jeroen de Man. An intelligent system for aggression de-escalation training. In *ECAI 2016*, pages 1805–1811. IOS Press, 2016.
- [52] Mihaela Botezatu, HÅkan Hult, Mesfin Kassaye Tessma, and Uno GH Fors. As time goes by: Stakeholder opinions on the implementation and use of a virtual patient simulation system. *Medical Teacher*, 32(11):e509–e516, 2010.
- [53] Jacqueline Bourdeau and Monique Grandbastien. Modeling tutoring knowledge. In *Advances in intelligent tutoring systems*, pages 123–143. Springer, 2010.
- [54] Mathieu Bourgeois, Patrick Taillandier, and Laurent Vercoouter. Ben: An agent architecture for explainable and expressive behavior in social simulation. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: First International Workshop, EXTRAAMAS 2019, Montreal, QC, Canada, May 13–14, 2019, Revised Selected Papers 1*, pages 147–163. Springer, 2019.
- [55] Hayet Brabra, Marcos Báez, Boualem Benatallah, Walid Gaaloul, Sara Bouguelia, and Shayan Zamanirad. Dialogue management in conversational systems: a review of approaches, challenges, and opportunities. *IEEE Transactions on Cognitive and Developmental Systems*, 2021.
- [56] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa. URL <https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa>.
- [57] Keith Brawner, Mike Hoffman, and Benjamin Nye. Architecture and ontology in the generalized intelligent framework for tutoring: 2018 update. In *7th Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium*, page 11, 2019.

- [58] Pete Bridge, Joanne Adeoye, Christopher N Edge, Vicky L Garner, Anne-Louise Humphreys, Sarah-Jane Ketterer, Joanne G Linforth, Anthony S Manning-Stanley, David Newsham, Denise Prescott, et al. Simulated placements as partial replacement of clinical training time: a delphi consensus study. *Clinical Simulation in Nursing*, 68:42–48, 2022.
- [59] Willem-Paul Brinkman, Dwi Hartanto, Ni Kang, Daniel de Vliegheer, Isabel L Kampmann, Nexhmedin Morina, Paul GM Emmelkamp, and Mark Neerincx. A virtual reality dialogue system for the treatment of social phobia. In *CHI'12 extended abstracts on human factors in computing systems*, pages 1099–1102. 2012.
- [60] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.
- [61] Susan M Brookhart. *How to give effective feedback to your students*. Ascd, 2017.
- [62] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [63] Emilia Brügge, Sarah Ricchizzi, Malin Arenbeck, Marius Niklas Keller, Lina Schur, Walter Stummer, Markus Holling, Max Hao Lu, and Dogus Darici. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Medical Education*, 24(1):1391, 2024.
- [64] Merijn Bruijnes. *Believable suspect agents: response and interpersonal style selection for an artificial suspect*. PhD thesis, University of Twente, 2016.
- [65] Merijn Bruijnes, Rieks op den Akker, Arno Hartholt, and Dirk Heylen. Virtual suspect william. In *Intelligent Virtual Agents: 15th International Conference, IVA 2015, Delft, The Netherlands, August 26-28, 2015, Proceedings 15*, pages 67–76. Springer, 2015.
- [66] Maike Bubeck and Wolfgang Bilsky. Value structure at an early age. *Swiss Journal of Psychology*, 63(1):31–41, 2004.
- [67] Paul-Christian Bürkner. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28, 2017.
- [68] David Lansing Cameron, Velibor Bobo Kovac, and May Olaug Horverak. Online chat-support for bullying prevention and follow-up: an analysis of the characteristics of chat-users and their experiences. *International Journal of Adolescence and Youth*, 28(1):2292060, 2023.
- [69] Francesco Cappa, Federica Rosso, and Darren Hayes. Monetary and social rewards for crowdsourcing. *Sustainability*, 11(10):2834, 2019.

- [70] Jeanne M Carey and Kelly Rossler. The how when why of high fidelity simulation. 2020.
- [71] David Carless. Longitudinal perspectives on students' experiences of feedback: A need for teacher–student partnerships. *Higher Education Research & Development*, 39(3):425–438, 2020.
- [72] Aishwarya Chandrasekaran, Rebecca Currano, Vafa Batool, Kaiping Chen, Elizabeth L Murnane, David Sirkin, and Matthew Louis Mauriello. Therapy for therapists: Design opportunities to support the psychological well-being of mental health workers. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–34, 2024.
- [73] Bo Chang. Reflection in learning. *Online learning*, 23(1):95–110, 2019.
- [74] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.
- [75] Alicja Chaszczewicz, Raj Shah, Ryan Louie, Bruce Arnow, Robert Kraut, and Diyi Yang. Multi-level feedback generation with large language models for empowering novice peer counselors. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4130–4161, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.227. URL <https://aclanthology.org/2024.acl-long.227/>.
- [76] Richard A Chechile. *Bayesian statistics for experimental scientists: a general introduction using distribution-free methods*. MIT Press, 2020.
- [77] Yuh-Ming Cheng, Lih-Shyang Chen, Hui-Chung Huang, Sheng-Feng Weng, Yong-Guo Chen, and Chyi-Her Lin. Building a general purpose pedagogical agent in a web-based multimedia clinical simulation system for medical education. *IEEE Transactions on Learning Technologies*, 2(3):216–225, 2009.
- [78] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.
- [79] Child Helpline International. Official website of child helpline international, 2025. URL <https://childhelplineinternational.org/>. Accessed: 15 December 2025.
- [80] Child Helpline International. Global Voices of Children and Young People, 2025. URL <https://childhelplineinternational.org/wp-content/uploads/2025/11/Global-Voices-of-Children-and-Young-People-2024-Factsheet-Final.pdf>.
- [81] Childline. About childline. <https://www.childline.org.uk/about/about-childline/>, 2025. Accessed: 2025-11-18.

- [82] Suzete Chiviawosky and Gabriele Wulf. Self-controlled feedback: does it enhance learning because performers get feedback when they need it? *Research quarterly for exercise and sport*, 73(4):408–415, 2002.
- [83] Suzete Chiviawosky and Gabriele Wulf. Self-controlled feedback is effective if it is based on the learner’s performance. *Research quarterly for exercise and sport*, 76(1): 42–48, 2005.
- [84] Mathieu Chollet, Stacy Marsella, and Stefan Scherer. Training public speaking with virtual social interactions: Effectiveness of real-time feedback and delayed feedback. *Journal on Multimodal User Interfaces*, 16(1):17–29, 2022.
- [85] Richard E Clark. Learning from serious games? arguments, evidence, and research suggestions. *Educational Technology*, 47(3):56–59, 2007.
- [86] Paul C Clements, Linda M Northrop, et al. *Software architecture: An executive overview*. Carnegie Mellon University, Software Engineering Institute, 1996.
- [87] CleVR, N.d. URL <https://clevr.net/> . Accessed 2023-09-11.
- [88] Jacob Cohen. A power primer. In Alan E. Kazdin, editor, *Methodological Issues and Strategies in Clinical Research*, pages 279–284. American Psychological Association, 2016. doi: 10.1037/14805-018.
- [89] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. Ethical judgment of agents’ behaviors in multi-agent systems. In *AAMAS*, pages 1106–1114, 2016.
- [90] David A Cook, Joshua Overgaard, V Shane Pankratz, Guilherme Del Fiol, and Chris A Aakre. Virtual patients using large language models: Scalable, contextualized simulation of clinician-patient dialogue with feedback. *Journal of Medical Internet Research*, 27:e68486, 2025.
- [91] Mark Core, H Chad Lane, and David Traum. Intelligent tutoring support for learners interacting with virtual humans. *Design Recommendations for Intelligent Tutoring Systems*, 2:249–257, 2014.
- [92] Scotty D Craig and Noah L Schroeder. Design principles for virtual humans in educational technology environments. In *Deep Comprehension*, pages 128–139. Routledge, 2018.
- [93] Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. No pizza for you: Value-based plan selection in bdi agents. In *IJCAI*, pages 178–184, 2017.
- [94] Carolyn Cyr and Peter W Dowrick. Burnout in crisisline volunteers. *Administration and policy in mental health and mental health services research*, 18(5):343–354, 1991.
- [95] Fred D Davis. *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. PhD thesis, Massachusetts Institute of Technology, 1985.

- [96] Anique de Beyn. *In gesprek met kinderen: de methodiek van de kindertelefoon*. NIZW, 2003.
- [97] De Kindertelefoon. *Jaarrekening 2023 Stichting de Kindertelefoon*. 2023. URL <https://jaarverslag.kindertelefoon.nl/resources/jaarverslagen/Jaarrekening%202023%20Stichting%20de%20Kindertelefoon.pdf>.
- [98] Edward L Deci and Richard M Ryan. *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media, 2013.
- [99] Orianna Demasi, Yu Li, and Zhou Yu. A multi-persona chatbot for hotline counselor training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3623–3636, 2020.
- [100] DialogueTrainer, N.d. URL <https://www.dialoguetrainer.com>. Accessed 2023-09-11.
- [101] Joao Dias, Samuel Mascarenhas, and Ana Paiva. Fatima modular: Towards an agent architecture with a generic appraisal framework. In *Emotion modeling*, pages 44–56. Springer, 2014.
- [102] Ding Ding. *Design and evaluation of simulated reflective thoughts in virtual reality exposure training*. PhD thesis, Delft University of Technology, 2020.
- [103] Ding Ding, Pascal Remeijnsen, Zian Song, Mark A Neerincx, and Willem-Paul Brinkman. Technology-supported social skills training systems: A systematic literature review. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 3249–3256. IEEE, 2024.
- [104] Dorian Doberstein, Nadja Agreiter, Marco Bäumer, Menglu Cui, Shaghayegh Abdollahzadegan, Diba Heidari, Nan Jiang, Markus Mentzel, Huangpan Zhang, Hao Zheng, et al. Cucomag-group reflection support in role-playing environments. *DeLFI 2016–Die 14. E-Learning Fachtagung Informatik*, 2016.
- [105] Anna K Döring, Andrea Blauensteiner, Katrin Aryus, Lisa Drögekamp, and Wolfgang Bilsky. Assessing values at an early age: The picture-based value survey for children (pbvs-c). *Journal of personality assessment*, 92(5):439–448, 2010.
- [106] Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos one*, 18(3):e0279720, 2023.
- [107] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 9, pages 48–59, 2021.
- [108] Lucile Dupuy, Jean-Arthur Micoulaud-Franchi, H el ene Cassoude‐salle, Orlane Ballot, Patrick Dehail, Bruno Aouizerate, Emmanuel Cuny, Etienne de Sevin, and Pierre Philip. Evaluation of a virtual agent to train medical students conducting psychiatric interviews for diagnosing major depressive disorders. *Journal of Affective Disorders*, 263:1–8, 2020.

- [109] Jacquelynne S Eccles and Allan Wigfield. Motivational beliefs, values, and goals. *Annual review of psychology*, 53(1):109–132, 2002.
- [110] Elsa Eiriksdottir and Richard Catrambone. Procedural instructions, principles, and examples: How to structure instructions for procedural tasks to enhance performance, learning, and transfer. *Human factors*, 53(6):749–770, 2011.
- [111] Katharina Emmerich, Katja Neuwald, Julia Othlinghaus, Sabrina Ziebarth, and H Ulrich Hoppe. Training conflict management in a collaborative virtual environment. In *International Conference on Collaboration and Technology*, pages 17–32. Springer, 2012.
- [112] Mica R Endsley. Situation awareness global assessment technique (sagat). In *Proceedings of the IEEE 1988 national aerospace and electronics conference*, pages 789–795. IEEE, 1988.
- [113] Peggy A Ertmer and Timothy J Newby. Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance improvement quarterly*, 6(4):50–72, 1993.
- [114] Irene Esteban-Millat, Francisco J Martínez-López, Maria Pujol-Jover, Juan Carlos Gázquez-Abad, and Alejandro Alegret. An extension of the technology acceptance model for online learning environments. *Interactive Learning Environments*, 26(7): 895–910, 2018.
- [115] European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union L 2024/1689, 2024. URL <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>. Published 12 July 2024; entered into force 1 August 2024.
- [116] Jacob K Farnsworth and Jennifer L Callahan. A model for addressing client–clinician value conflict. *Training and Education in Professional Psychology*, 7(3):205, 2013.
- [117] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies*, 132:138–161, 2019.
- [118] Dan Feng, Pedro Sequeira, Elin Carstensdottir, Magy Seif El-Nasr, and Stacy Marsella. Learning generative models of social interactions with humans-in-the-loop. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 509–516. IEEE, 2018.
- [119] Leon Festinger. Cognitive dissonance. *Scientific American*, 207(4):93–106, 1962.
- [120] Elisabeth Fischer and Martin Hänze. How do university teachers’ values and beliefs affect their teaching? *Educational Psychology*, 40(3):296–317, 2020.
- [121] Kurt W Fischer. A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological review*, 87(6):477, 1980.

- [122] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, and Willem-Paul Brinkman. Questionnaire items for evaluating artificial social agents-expert generated, content validated and reliability analysed. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 84–86, 2021.
- [123] Siska Fitrianie, Corine Horsch, Robbert Jan Beun, Fiemke Griffioen-Both, and Willem-Paul Brinkman. 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, 2021.
- [124] Siska Fitrianie, Merijn Bruijnes, Fengxiang Li, Amal Abdulrahman, and Willem-Paul Brinkman. The artificial-social-agent questionnaire: establishing the long and short questionnaire versions. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2022.
- [125] Siska Fitrianie, Merijn Bruijnes, Amal Abdulrahman, and Willem-Paul Brinkman. 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*, page 103482, 2025.
- [126] Rowanne Fleck and Geraldine Fitzpatrick. Reflecting on reflection: framing a design landscape. In *Proceedings of the 22nd conference of the computer-human interaction special interest group of australia on computer-human interaction*, pages 216–223, 2010.
- [127] Kristin Fraser, Bruce Wright, Louis Girard, Janet Tworek, Mike Paget, Lisa Welikovich, and Kevin McLaughlin. Simulation training improves diagnostic performance on a real patient with similar clinical findings. *Chest*, 139(2):376–381, 2011.
- [128] Samuel Kernan Freire, Chaofan Wang, and Evangelos Niforatos. Conversational assistants in knowledge-intensive contexts: An evaluation of llm-versus intent-based systems. *arXiv preprint arXiv:2402.04955*, 2024.
- [129] Laurent Frering, Gerald Steinbauer-Wagner, and Andreas Holzinger. Integrating belief-desire-intention agents with large language models for reliable human–robot interaction and explainable artificial intelligence. *Engineering Applications of Artificial Intelligence*, 141:109771, 2025.
- [130] Thomas C Froehle, Sharon E Robinson, and Wayne J De Kurpius. Enhancing the effects of modeling through role-play practice. *Counselor Education and Supervision*, 22(3):197–206, 1983.
- [131] Ruben Georges Fukkink, Suzan Bruns, and Rudy Ligtvoet. Voices of children from around the globe; an international analysis of children’s issues at child helplines. *Children & Society*, 30(6):510–519, 2016.
- [132] David C Funder and Daniel J Ozer. Evaluating effect size in psychological research: Sense and nonsense. *Advances in methods and practices in psychological science*, 2(2): 156–168, 2019.

- [133] Robert Mills Gagné. *The conditions of learning and theory of instruction*. Holt, 1985.
- [134] Anthony G Gallagher, E Matt Ritter, Howard Champion, Gerald Higgins, Marvin P Fried, Gerald Moses, C Daniel Smith, and Richard M Satava. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. *Annals of surgery*, 241(2):364, 2005.
- [135] Patrick Gebhard, Tanja Schneeberger, Elisabeth André, Tobias Baur, Ionut Damian, Gregor Mehlmann, Cornelius König, and Markus Langer. Serious games for training social skills in job interviews. *IEEE Transactions on Games*, 11(4):340–351, 2018.
- [136] Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. The belief-desire-intention model of agency. In *International workshop on agent theories, architectures, and languages*, pages 1–10. Springer, 1998.
- [137] Michael Georgeff, Barney Pell, Martha Pollack, Milind Tambe, and Michael Wooldridge. The belief-desire-intention model of agency. In *Intelligent Agents V: Agents Theories, Architectures, and Languages: 5th International Workshop, ATAL'98 Paris, France, July 4–7, 1998 Proceedings 5*, pages 1–10. Springer, 1999.
- [138] Kallirroi Georgila, Mark G Core, Benjamin D Nye, Shamyia Karumbaiah, Daniel Auerbach, and Maya Ram. Using reinforcement learning to optimize the policies of an intelligent tutoring system for interpersonal skills training. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 737–745, 2019.
- [139] Rachele K Gould, Thais Moreno Soares, Paola Arias-Arévalo, Mariana Cantú-Fernandez, Dana Baker, Harold N Eyster, Rain Kwon, Lauren Prox, Julian Rode, Andres Suarez, et al. The role of value (s) in theories of human behavior. *Current Opinion in Environmental Sustainability*, 64:101355, 2023.
- [140] Arthur C Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, and Max M Louwerse. Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36:180–192, 2004.
- [141] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier, 2013.
- [142] Andrina Granić and Nikola Marangunić. Technology acceptance model in educational context: A systematic literature review. *British Journal of Educational Technology*, 50(5):2572–2593, 2019.
- [143] Jonathan Gratch, David DeVault, and Gale Lucas. The benefits of virtual humans for teaching negotiation. In *International Conference on Intelligent Virtual Agents*, pages 283–294. Springer, 2016.

- [144] Sharon Grundmann. A bdi-based virtual agent for training child helpline counsellors. 2022. URL <https://repository.tudelft.nl/islandora/object/uuid%3Af04f8f0b-9ab9-4f1c-a19c-43b164d45cce>.
- [145] Sharon Grundmann, Mohammed Al Owayyed, Merijn Bruijnes, Ellen Vroonhof, and Willem-Paul Brinkman. Lilobot: A cognitive conversational agent to train counsellors at children’s helplines: Design and initial evaluation. *Journal of Medical Systems*, 49(1):5, 2025.
- [146] Timothy C Guetterman, Rae Sakakibara, Srikar Baireddy, Frederick W Kron, Mark W Scerbo, James F Cleary, and Michael D Fetters. 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, 2019.
- [147] Önder Gürçan. Llm-augmented agent-based modelling for social simulations: Challenges and opportunities. *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 134–144, 2024.
- [148] Onder Gurcan. Llm-augmented agent-based modelling for social simulations: Challenges and opportunities, 2024. URL <https://arxiv.org/abs/2405.06700>.
- [149] Mark Halstead and Monica J Taylor. *Values in education and education in values*. Routledge, 2005.
- [150] Paul HP Hanel, Gregory R Maio, Ana KS Soares, Katia C Vione, Gabriel L de Holanda Coelho, Valdiney V Gouveia, Appasaheb C Patil, Shanmukh V Kamble, and Antony SR Manstead. Cross-cultural differences and similarities in human value instantiation. *Frontiers in psychology*, 9:849, 2018.
- [151] Maaïke Harbers, Karel Van Den Bosch, and John-Jules Ch Meyer. A study into preferred explanations of virtual agent behavior. In *International Workshop on Intelligent Virtual Agents*, pages 132–145. Springer, 2009.
- [152] Jan-Gerrit Harms, Pavel Kucherbaev, Alessandro Bozzon, and Geert-Jan Houben. Approaches for dialog management in conversational agents. *IEEE Internet Computing*, 23(2):13–22, 2018.
- [153] Dwi Hartanto, Isabel L Kampmann, Nexhmedin Morina, Paul GM Emmelkamp, Mark A Neerinx, and Willem-Paul Brinkman. Controlling social stress in virtual reality environments. *PloS one*, 9(3):e92804, 2014.
- [154] Arno Hartholt and Sharon Mozgai. Platforms and tools for sia research and development. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application*, pages 261–304. 2022.
- [155] Arno Hartholt, David Traum, Stacy C Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch. All together now, introducing

- the virtual human toolkit. In *International Workshop on Intelligent Virtual Agents*, pages 368–381. Springer, 2013.
- [156] Arno Hartholt, Ed Fast, Zongjian Li, Kevin Kim, Andrew Leeds, and Sharon Mozgai. Re-architecting the virtual human toolkit: towards an interoperable platform for embodied conversational agent research and development. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2022.
- [157] Andreas Hartmann, Joanne Vinke-de Kruijf, and Ruben Van Weesep. Asking the right questions: The role of reflection for learning in and between projects. *International Journal of Project Management*, 41(5):102494, 2023.
- [158] John Hattie and Helen Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.
- [159] Neville Hatton and David Smith. Reflection in teacher education: Towards definition and implementation. *Teaching and teacher education*, 11(1):33–49, 1995.
- [160] Matthew Jensen Hays, Julia C Campbell, Matthew A Trimmer, Joshua C Poore, Andrea K Webb, and Teresa K King. Can role-play with virtual humans teach interpersonal skills. In *Interservice/Industry Training, Simulation and Education Conference (IITSEC)*, 2012.
- [161] Samaneh Heidari, Maarten Jensen, and Frank Dignum. Simulations with values. In *Advances in Social Simulation: Looking in the Mirror*, pages 201–215. Springer, 2020.
- [162] Michael V Heinz, Daniel M Mackin, Brianna M Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A Banta, Abi D Jewett, Abigail J Salzhauer, Tess Z Griffin, and Nicholas C Jacobson. Randomized trial of a generative ai chatbot for mental health treatment. *NEJM AI*, 2(4):A1oa2400802, 2025.
- [163] Jennifer Hettema, Julie Steele, and William R Miller. Motivational interviewing. *Annu. Rev. Clin. Psychol.*, 1(1):91–111, 2005.
- [164] Randall W Hill Jr, Jonathan Gratch, Stacy Marsella, Jeff Rickel, William R Swartout, and David R Traum. Virtual humans in the mission rehearsal exercise system. *Künstliche Intell.*, 17(4):5, 2003.
- [165] Dennis E Hinkle, William Wiersma, Stephen G Jurs, et al. *Applied statistics for the behavioral sciences*, volume 663. Houghton Mifflin Boston, 2003.
- [166] Atsusi Hirumi, Teresa Johnson, Ramsamooj Javier Reyes, Benjamin Lok, Kyle Johnsen, Diego J Rivera-Gutierrez, Kenneth Bogert, Stacey Kubovec, Michael Eakins, Andrea Kleinsmith, et al. Advancing virtual patient simulations through design research and inter play: part ii—integration and field test. *Educational technology research and development*, 64:1301–1335, 2016.
- [167] Atsusi Hirumi, Andrea Kleinsmith, Kyle Johnsen, Stacey Kubovec, Michael Eakins, Kenneth Bogert, Diego J Rivera-Gutierrez, Ramsamooj Javier Reyes, Benjamin Lok, and Juan Cendan. Advancing virtual patient simulations through design research

- and interplay: part i: design and development. *Educational Technology Research and Development*, 64:763–785, 2016.
- [168] April Hoang, Stevie-Jae Hepburn, Alina Morawska, and Matthew R Sanders. The effect of self-reflection on the outcomes of online clinical skills training: a comparative study. *Advances in Health Sciences Education*, pages 1–19, 2025.
- [169] Friederike Holderried, Christian Stegemann-Philipps, Lea Herschbach, Julia-Astrid Moldt, Andrew Nevins, Jan Griewatz, Martin Holderried, Anne Herrmann-Werner, Teresa Festl-Wietek, Moritz Mahling, et al. 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, 2024.
- [170] Wayne Holmes and Kaśka Porayska-Pomsta. The ethics of artificial intelligence in education. *Lontoo: Routledge*, pages 621–653, 2023.
- [171] Mohammed Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 697–706, 2013.
- [172] Adam O Horvath and Leslie S Greenberg. Development and validation of the working alliance inventory. *Journal of counseling psychology*, 36(2):223, 1989.
- [173] Rui Huang, Albert D Ritzhaupt, Max Sommer, Jiawen Zhu, Anita Stephen, Natercia Valle, John Hampton, and Jingwei Li. The impact of gamification in educational settings on student learning outcomes: A meta-analysis. *Educational Technology Research and Development*, 68(4):1875–1901, 2020.
- [174] Eva Hudlicka. Virtual affective agents and therapeutic games. In *Artificial intelligence in behavioral and mental health care*, pages 81–115. Elsevier, 2016.
- [175] Lurdes YT Inoue, Donald A Berry, and Giovanni Parmigiani. Relationship between bayesian and frequentist sample size determination. *The American Statistician*, 59(1): 79–87, 2005.
- [176] Child Helpline International. Voices of children & young people around the world: Global child helpline data from 2023, 2024. URL <https://childhelplineinternational.org/voices-2023-data/>. Accessed: 2025-04-01.
- [177] Allen E Ivey, Mary Bradford Ivey, and Lynn Simek-Downing. *Counseling and psychotherapy: Integrating skills, theory, and practice*. Prentice-Hall, Inc, 1987.
- [178] Brian W Jack, Timothy Bickmore, Leanne Yinusa-Nyahkoon, Matthew Reichert, Clevanne Julce, Nireesha Sidduri, Jessica Martin-Howard, Zhe Zhang, Elisabeth Woodhams, Juan Fernandez, et al. 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, 2020.

- [179] Simon Jacklin, Neal Maskrey, Stephen Chapman, et al. Improving shared decision making between patients and clinicians: design and development of a virtual patient simulation tool. *JMIR medical education*, 4(2):e10088, 2018.
- [180] Simon Jacklin, Stephen Chapman, and Neal Maskrey. Virtual patient educational intervention for the development of shared decision-making skills: a pilot study. *BMJ Simulation & Technology Enhanced Learning*, 5(4):215, 2019.
- [181] David Jackson and Annabel Latham. Talk to the ghost: The storybox methodology for faster development of storytelling chatbots. *Expert Systems with Applications*, 190:116223, 2022.
- [182] Wilco Janssen, Jeroen van Raak, Yannick van der Lucht, Wouter van Ballegooijen, and Saskia Mérelle. 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, 2022.
- [183] Simone Jennissen, Stefan Schumacher, Diana Rucli, Melinda Hal, András Székely, Derek De Beurs, and Ulrike Dinger. Competency-based training and assessment of listening skills: a waitlist-controlled study in european telephone emergency services. *Clinical psychology in Europe*, 4(4):e7933, 2022.
- [184] Johan Jeuring, Frans Grosfeld, Bastiaan Heeren, Michiel Hulsbergen, Richta IJntema, Vincent Jonker, Nicole Mastenbroek, Maarten van der Smagt, Frank Wijmans, Manjanne Wolters, et al. Communicate!—a serious game for communication skills—. In *Design for Teaching and Learning in a Networked World: 10th European Conference on Technology Enhanced Learning, EC-TEL 2015, Toledo, Spain, September 15-18, 2015, Proceedings 10*, pages 513–517. Springer, 2015.
- [185] Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J Su, Camillo J Taylor, and Dan Roth. A peek into token bias: Large language models are not yet genuine reasoners. *arXiv preprint arXiv:2406.11050*, 2024.
- [186] YA Jimenez, F Gray, L Di Michele, S Said, W Reed, and P Kench. Can simulation-based education or other education interventions replace clinical placement in medical radiation sciences? a narrative review. *Radiography*, 29(2):421–427, 2023.
- [187] Hui Jin, Jamie N Mikeska, Hayat Hokayem, and Elia Mavronikolas. Toward coherence in curriculum, instruction, and assessment: A review of learning progression literature. *Science Education*, 103(5):1206–1234, 2019.
- [188] Emmanuel Johnson, Gale Lucas, Peter Kim, and Jonathan Gratch. Intelligent tutoring system for negotiation skills training. In *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part II 20*, pages 122–127. Springer, 2019.
- [189] Jeff Johnson and Austin Henderson. Conceptual models: begin by designing what to design. *interactions*, 9(1):25–32, 2002.

- [190] David H Jonassen and Philip Henning. Mental models: Knowledge in the head and knowledge in the world. *Educational technology*, pages 37–42, 1999.
- [191] Sean HK Kang. Spaced repetition promotes efficient and effective learning: Policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences*, 3(1):12–19, 2016.
- [192] Nikitas Karanikolas, Eirini Manga, Nikoletta Samaridi, Eleni Tousidou, and Michael Vassilakopoulos. Large language models versus natural language understanding and generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, pages 278–290, 2023.
- [193] Jeffrey D Karpicke and Henry L Roediger III. Repeated retrieval during learning is the key to long-term retention. *Journal of memory and language*, 57(2):151–162, 2007.
- [194] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.
- [195] Tessa ML Kaufman, Tina Kretschmer, Gijs Huitsing, and René Veenstra. Why does a universal anti-bullying program not help all children? explaining persistent victimization during an intervention. *Prevention science*, 19(6):822–832, 2018.
- [196] Ralph L Keeney. Value-driven expert systems for decision support. *Decision support systems*, 4(4):405–412, 1988.
- [197] Ralph L Keeney. Value-focused thinking: Identifying decision opportunities and creating alternatives. *European Journal of operational research*, 92(3):537–549, 1996.
- [198] Patrick Kenny, Arno Hartholt, Jonathan Gratch, William Swartout, David Traum, Stacy Marsella, and Diane Piepol. Building interactive virtual humans for training environments. In *Proceedings of i/itsec*, volume 174, pages 911–916, 2007.
- [199] Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Yi-Shan Tsai, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Dragan Gašević. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3:100074, 2022.
- [200] De Kindertelefoon. De kindertelefoon - always someone to talk to. 2023. URL <https://www.kindertelefoon.nl/resources/pdf-bestanden/cl-en-2023.pdf>.
- [201] De Kindertelefoon. De kindertelefoon. 2024. URL <https://jaarverslag.kindertelefoon.nl/resources/jaarverslagen/Jaarrekening%202024%20Stichting%20De%20Kindertelefoon.pdf>.

- [202] James D Kirkpatrick and Wendy Kayser Kirkpatrick. *Kirkpatrick's four levels of training evaluation*. Association for Talent Development, 2016.
- [203] Howard Kirschenbaum. *Values clarification in counseling and psychotherapy: Practical strategies for individual and group settings*. Oxford University Press, 2013.
- [204] Paul Kirschner, John Sweller, and Richard E Clark. Why unguided learning does not work: An analysis of the failure of discovery learning, problem-based learning, experiential learning and inquiry-based learning. *Educational psychologist*, 41(2): 75–86, 2006.
- [205] Karen Strohm Kitchner. Cognition, metacognition, and epistemic cognition: A three-level model of cognitive processing. *Human development*, 26(4):222–232, 1983.
- [206] Andrea Kleinsmith, Diego Rivera-Gutierrez, Glen Finney, Juan Cendan, and Benjamin Lok. Understanding empathy training with virtual patients. *Computers in human behavior*, 52:151–158, 2015.
- [207] Jennifer K Knight, Daniel C Weaver, Melanie E Peffer, and Zachary S Hazlett. Relationships between prediction accuracy, metacognitive reflection, and performance in introductory genetics students. *CBE—Life Sciences Education*, 21(3):ar45, 2022.
- [208] Ekaterina Kochmar, Dung Do Vu, Robert Belfer, Varun Gupta, Iulian Vlad Serban, and Joelle Pineau. Automated data-driven generation of personalized pedagogical interventions in intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 32(2):323–349, 2022.
- [209] Kognito, N.d. URL <https://kognito.com/> . Accessed 2023-09-11.
- [210] David A Kolb. *Experiential learning: Experience as the source of learning and development*. FT press, 2014.
- [211] Christine M Korsgaard. *The sources of normativity*. Cambridge University Press, 1996.
- [212] Iuliia Kotseruba and John K Tsotsos. 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, 53(1): 17–94, 2020.
- [213] Frederick W Kron, Michael D Fetters, Mark W Scerbo, Casey B White, Monica L Lypton, Miguel A Padilla, Gayle A Gliva-McConvey, Lee A Belfore II, Temple West, Amelia M Wallace, et al. Using a computer simulation for teaching communication skills: A blinded multisite mixed methods randomized controlled trial. *Patient education and counseling*, 100(4):748–759, 2017.
- [214] John K Kruschke. Rejecting or accepting parameter values in bayesian estimation. *Advances in methods and practices in psychological science*, 1(2):270–280, 2018.
- [215] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. Supporting the demand on mental health services with ai-based conversational large language models (llms). *BioMedInformatics*, 4(1):8–33, 2023.

- [216] John E Laird. *The Soar cognitive architecture*. MIT press, 2019.
- [217] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.
- [218] H Chad Lane, Mark G Core, David Gomboc, Arshish Karnavat, and Milton Rosenberg. Intelligent tutoring for interpersonal and intercultural skills. In *Interservice/Industry Training, Simulation and Education Conference (IITSEC)*, volume 111, 2007.
- [219] Douglas P Larsen, Daniel A London, and Amanda R Emke. Using reflection to influence practice: student perceptions of daily reflection in clinical education. *Perspectives on medical education*, 5(5):285–291, 2016.
- [220] Lisa M Larson and Jeffrey A Daniels. Review of the counseling self-efficacy literature. *The Counseling Psychologist*, 26(2):179–218, 1998.
- [221] Alyssa P Lawson, Richard E Mayer, Nicoletta Adamo-Villani, Bedrich Benes, Xingyu Lei, and Justin Cheng. Do learners recognize and relate to the emotions displayed by virtual instructors? *International Journal of Artificial Intelligence in Education*, 31: 134–153, 2021.
- [222] Jihyun Lee, Hyungsin Kim, Kwan Hoon Kim, Daeun Jung, Tanisha Jowsey, and Craig S Webster. Effective virtual patient simulators for medical communication training: a systematic review. *Medical education*, 54(9):786–795, 2020.
- [223] Jina Lee, David DeVault, David Traum, and Stacy Marsella. Thoughts on fml: Behavior generation in the virtual human communication architecture. *Proceedings of FML*, pages 83–95, 2008.
- [224] June Lee. Effectiveness of computer-based instructional simulation: a meta analysis. *International journal of instructional media*, 26(1):71, 1999.
- [225] Stephen Lee-Cheong, Mariam Alaverdashvili, Mackenzie Jardine, Vidhi Shivani Maharaj, and Nathan Kolla. Burnout and professional quality of life amongst crisis hotline responders: A cross-sectional survey in canada during covid-19. In *Healthcare*, volume 13, page 1025. MDPI, 2025.
- [226] Alan M Leslie. Tomm, toby, and agency: Core architecture and domain specificity. *Mapping the mind: Domain specificity in cognition and culture*, 29:119–48, 1994.
- [227] Magdeleine DN Lew and Henk G Schmidt. Self-reflection and academic performance: is there a relationship? *Advances in Health Sciences Education*, 16(4):529–545, 2011.
- [228] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- [229] Ming-Hung Lin, Huang-Cheng Chen, and Kuang-Sheng Liu. 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, 2017.

- [230] Yun Ling, Harold T Nefs, Willem-Paul Brinkman, Chao Qu, and Ingrid Heynderickx. The effect of perspective on presence and space perception. *PloS one*, 8(11):e78513, 2013.
- [231] Enrico Liscio, Luciano C Siebert, Catholijn M Jonker, and Pradeep K Murukannaiah. Value preferences estimation and disambiguation in hybrid participatory systems. *Journal of Artificial Intelligence Research*, 82:819–850, 2025.
- [232] STEPHEN N LITTLEFIELD and SONDRRA Z KOFF. Hotlines that burnout: A study of the factors which contribute to the failure of crisis intervention centers. *Journal of health and human resources administration*, pages 279–295, 1986.
- [233] Chenxiao Liu, Zheyong Xie, Sirui Zhao, Jin Zhou, Tong Xu, Minglei Li, and Enhong Chen. Speak from heart: an emotion-guided llm-based multimodal method for emotional dialogue generation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 533–542, 2024.
- [234] Jose Llanes-Jurado, Lucía Gómez-Zaragozá, Maria Eleonora Minissi, Mariano Alcañiz, and Javier Marín-Morales. Developing conversational virtual humans for social emotion elicitation based on large language models. *Expert Systems with Applications*, 246:123261, 2024.
- [235] Michael Loizou, Sylvester Arnab, Petros Lameris, Thomas Hartley, Fernando Loizides, Praveen Kumar, and Dana Sumilo. 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, 2024.
- [236] Do Xuan Long, Hai Nguyen Ngoc, Tiviatis Sim, Hieu Dao, Shafiq Joty, Kenji Kawaguchi, Nancy F Chen, and Min-Yen Kan. Llms are biased towards output formats! systematically evaluating and mitigating output format bias of llms. *arXiv preprint arXiv:2408.08656*, 2024.
- [237] Francisca Losavio, Ledis Chirinos, and Maria A Pérez. Quality models to design software architectures. In *Proceedings Technology of Object-Oriented Languages and Systems. TOOLS 38*, pages 123–135. IEEE, 2001.
- [238] Sasha Luccioni, Bruna Trevelin, and Margaret Mitchell. The environmental impacts of ai—primer. *Hugging Face Blog*, 2024.
- [239] Kellen Maicher, Douglas Danforth, Alan Price, Laura Zimmerman, Bruce Wilcox, Beth Liston, Holly Cronau, Laurie Belknap, Cynthia Ledford, David Way, et al. Developing a conversational virtual standardized patient to enable students to practice history-taking skills. *Simulation in Healthcare*, 12(2):124–131, 2017.
- [240] Henry W Maier. Role playing: structures and educational objectives. *The International Child and Youth Care Network*, 36, 2002.
- [241] Dominique Makowski, Mattan S Ben-Shachar, SH Annabel Chen, and Daniel Lüdecke. Indices of effect existence and significance in the bayesian framework. *Frontiers in psychology*, 10:2767, 2019.

- [242] Dominique Makowski, Mattan S Ben-Shachar, and Daniel Lüdecke. bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of open source software*, 4(40):1541, 2019.
- [243] Brent Mallinckrodt. Attachment, social competencies, social support, and interpersonal process in psychotherapy. *Psychotherapy research*, 10(3):239–266, 2000.
- [244] Zainab Maqsood, Madiha Sajjad, and Raheela Yasmin. Effect of feedback-integrated reflection, on deep learning of undergraduate medical students in a clinical setting. *BMC Medical Education*, 25(1):66, 2025.
- [245] Stacy C Marsella and Jonathan Gratch. Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90, 2009.
- [246] Andrea Martone and Stephen G Sireci. Evaluating alignment between curriculum, assessment, and instruction. *Review of educational research*, 79(4):1332–1361, 2009.
- [247] Samuel Mascarenhas, Joao Dias, Rui Prada, and Ana Paiva. A dimensional model for cultural behavior in virtual agents. *Applied Artificial Intelligence*, 24(6):552–574, 2010.
- [248] Samuel Mascarenhas, Manuel Guimarães, Rui Prada, Pedro A Santos, João Dias, and Ana Paiva. Fatima toolkit: Toward an accessible tool for the development of socio-emotional agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 12(1):1–30, 2022.
- [249] Sharna L Mathieu, Riaz Uddin, Morgan Brady, Samantha Batchelor, Victoria Ross, Susan H Spence, David Watling, and Kairi Kõlves. Systematic review: The state of research into youth helplines. *Journal of the American Academy of Child & Adolescent Psychiatry*, 60(10):1190–1233, 2021.
- [250] Richard E Mayer. Should there be a three-strikes rule against pure discovery learning? *American psychologist*, 59(1):14, 2004.
- [251] John McLeod. *An introduction to counselling*. McGraw-hill education (UK), 2013.
- [252] Graham McPhail. The search for deep learning: A curriculum coherence model. *Journal of Curriculum Studies*, 53(4):420–434, 2021.
- [253] Marjorie McShane, Sergei Nirenburg, Bruce Jarrell, Stephen Beale, and George Fantry. Maryland virtual patient: A knowledge-based, language-enabled simulation and training system. *Bio Algorithms Med Syst.*, 5(9):57–63, 2009.
- [254] Michael McTear, Zoraida Callejas, and David Griol. Creating a conversational interface using chatbot technology. In *The conversational interface*, pages 125–159. Springer, 2016.
- [255] Michael Frederick McTear, Zoraida Callejas, and David Griol. *The conversational interface*, volume 6. Springer, 2016.

- [256] Johnathan Mell and Jonathan Gratch. Grumpy & pinocchio: answering human-agent negotiation questions through realistic agent design. In *Proceedings of the 16th conference on autonomous agents and multiagent systems*, pages 401–409, 2017.
- [257] Ersilia Menesini, Annalaura Nocentini, and Marina Camodeca. Morality, values, traditional bullying, and cyberbullying in adolescence. *British Journal of Developmental Psychology*, 31(1):1–14, 2013.
- [258] José Mercado, Ismael Edrein Espinosa-Curiel, and Juan Martínez-Miranda. Embodied conversational agents providing motivational interviewing to improve health-related behaviors: Scoping review. *Journal of medical Internet research*, 25:e52097, 2023.
- [259] M David Merrill. *First principles of instruction*. John Wiley & Sons, 2012.
- [260] William R Miller and Kathy A Mount. A small study of training in motivational interviewing: Does one workshop change clinician and client behavior? *Behavioural and cognitive psychotherapy*, 29(4):457–471, 2001.
- [261] William R Miller and Stephen Rollnick. *Motivational interviewing: Helping people change*. Guilford press, 2012.
- [262] William R Miller, Carolina E Yahne, Theresa B Moyers, James Martinez, and Matthew Pirritano. A randomized trial of methods to help clinicians learn motivational interviewing. *Journal of consulting and Clinical Psychology*, 72(6):1050, 2004.
- [263] Valérie Millette and Marylene Gagné. Designing volunteers’ tasks to maximize motivation, satisfaction and performance: The impact of job characteristics on volunteer engagement. *Motivation and emotion*, 32(1):11–22, 2008.
- [264] Chinmaya Mishra, Rinus Verdonshot, Peter Hagoort, and Gabriel Skantze. Real-time emotion generation in human-robot dialogue using large language models. *Frontiers in Robotics and AI*, 10:1271610, 2023.
- [265] Jennifer A Moon. *Reflection in learning and professional development: Theory and practice*. Routledge, 2013.
- [266] Richard D. Morey and Jeffrey N. Rouder. *Bayesfactor: Computation of Bayes Factors for Common Designs*. 2018. URL <https://CRAN.R-project.org/package=BayesFactor>.
- [267] Connie M Moss and Susan M Brookhart. *Learning targets: Helping students aim for understanding in today’s lesson*. ASCD, 2012.
- [268] Quim Motger, Xavier Franch, and Jordi Marco. Software-based dialogue systems: Survey, taxonomy and challenges. *ACM Computing Surveys (CSUR)*, 2022.
- [269] Maya Mouawad and Brian H Kleiner. New developments in customer service training. *Managing Service Quality: An International Journal*, 6(2):49–56, 1996.

- [270] Jase Moussa-Inaty. Reflective writing through the use of guiding questions. *International Journal of Teaching and Learning in Higher Education*, 27(1):104–113, 2015.
- [271] Keith Muller. *Statistical power analysis for the behavioral sciences*. Taylor & Francis, 1989.
- [272] Tijmen Muller, Annerieke Heuvelink, Karel van den Bosch, and Ivo Swartjes. Glen-garry glen ross: Using bdi for sales game dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 8, pages 167–172, 2012.
- [273] Prasanth Murali, Ha Trinh, Lazlo Ring, and Timothy Bickmore. A friendly face in the crowd: Reducing public speaking anxiety with an emotional support agent in the audience. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 156–163, 2021.
- [274] Prasanth Murali, Farnaz Nouraei, Mina Fallah, Aisling Kearns, Keith Rebello, Teresa O’Leary, Rebecca Perkins, Natalie Pierre Joseph, Julien Dedier, Michael Paasche-Orlow, et al. Training lay counselors with virtual agents to promote vaccination. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2022.
- [275] Mursion, N.d. URL <https://www.mursion.com/> . Accessed 2023-09-11.
- [276] Sylvie Naar and Mariann Suarez. *Motivational interviewing with adolescents and young adults*. Guilford Publications, 2021.
- [277] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.
- [278] Tal Nakash, Tom Haller, Maya Shekel, Dan Pollak, Moti Lewenchuse, Anat Brunstein Klomek, and Doron Friedman. Increasing resilience and preventing suicide: training and interventions with a distressed virtual human in virtual reality. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2022.
- [279] National Society for the Prevention of Cruelty to Children (NSPCC). Childline – volunteer with nspcc. <https://join-us.nspcc.org.uk/volunteers/volunteers/childline/>. Accessed: 2025-10-22.
- [280] Zeynab Nazirova and Simonovits Borbala. Values, attitudes and the behaviour paradigm: a systematic literature review. *Journal of Human Values*, 30(2):214–239, 2024.
- [281] Mark A Neerincx, Jasper van der Waa, Frank Kaptein, and Jurriaan van Diggelen. Using perceptual and cognitive explanations for enhanced human-agent team performance. In *International Conference on Engineering Psychology and Cognitive Ergonomics*, pages 204–214. Springer, 2018.

- [282] Richard P Niemiec, Melanie F Sikorski, and Herbert J Walberg. Comparing the cost-effectiveness of tutoring and computer-based instruction. *Journal of Educational Computing Research*, 5(4):395–407, 1989.
- [283] Farnaz Nouraei, Keith Rebello, Mina Fallah, Prasanth Murali, Haley Matuszak, Valerie Jap, Andrea Parker, Michael Paasche-Orlow, and Timothy Bickmore. Virtual agent-based communication skills training to facilitate health persuasion among peers. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–24, 2025.
- [284] NSPCC. Childline and nspcc helpline statistics, June 2025. URL <https://learning.nspcc.org.uk/research-resources/statistics-briefings/childline-nspcc-helpline-statistics>. Accessed: 2025-11-06.
- [285] J.C. Nunnally and I.H. Bernstein. *Psychometric theory*. McGraw–Hill, New York, 1994.
- [286] Hyacinth S Nwana. Intelligent tutoring systems: an overview. *Artificial Intelligence Review*, 4(4):251–277, 1990.
- [287] Magalie Ochs, Daniel Mestre, Grégoire De Montcheuil, Jean-Marie Pergandi, Jorane Saubesty, Evelyne Lombardo, Daniel Francon, and Philippe Blache. Training doctors’ social skills to break bad news: evaluation of the impact of virtual environment displays on the sense of presence. *Journal on Multimodal User Interfaces*, 13(1):41–51, 2019.
- [288] Pok-Ja Oh, Kyeong Deok Jeon, and Myung Suk Koh. The effects of simulation-based learning using standardized patients in nursing students: A meta-analysis. *Nurse education today*, 35(5):e6–e15, 2015.
- [289] Joost van Oijen, Willem van Doesburg, and Frank Dignum. Goal-based communication using bdi agents as virtual humans in training: An ontology driven dialogue system. In *International Workshop on Agents for Games and Simulations*, pages 38–52. Springer, 2010.
- [290] Suman Ojha, Jonathan Vitale, and Mary-Anne Williams. Computational emotion models: a thematic review. *International Journal of Social Robotics*, 13(6):1253–1279, 2021.
- [291] Stefan Olafsson, Teresa O’Leary, and Timothy Bickmore. 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*, pages 31–40, 2019.
- [292] Committee on Quality of Health Care in America. *Crossing the quality chasm: a new health system for the 21st century*. National Academies Press, 2001.

- [293] Candice Oster, Adrian Schoo, John Litt, Andrea Morello, Richard Leibbrandt, Christopher Antonello, David Powers, Belinda Lange, Anthony Maeder, and Sharon Lawn. 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, 2020.
- [294] Candice Oster, Richard Leibbrandt, Adrian Schoo, Belinda Lange, John Litt, Andrea Morello, Christopher Antonello, David Powers, Anthony Maeder, and Sharon Lawn. A feasibility study of teaching motivational interviewing in a fully online environment using a virtual client. *International Journal of Health Promotion and Education*, pages 1–12, 2022.
- [295] Julia Othlinghaus-Wulhorst, Anne Mainz, and H Ulrich Hoppe. Training customer complaint management in a virtual role-playing game: a user study. In *Transforming Learning with Meaningful Technologies: 14th European Conference on Technology Enhanced Learning, EC-TEL 2019, Delft, The Netherlands, September 16–19, 2019, Proceedings 14*, pages 436–449. Springer, 2019.
- [296] Kimberly H McManama O’Brien, Shai Fuxman, Laura Humm, Nicole Tirone, Warren Jay Pires, Andrea Cole, and Julie Goldstein Grumet. Suicide risk assessment training using an online virtual patient simulation. *Mhealth*, 5, 2019.
- [297] IGAL Parwata, I Nyoman Laba Jayanta, and I Wayan Widiana. Improving metacognitive ability and learning outcomes with problem-based revised bloom’s taxonomy oriented learning activities. *Emerging Science Journal*, 7(2):569–577, 2023.
- [298] CH Patterson. Values in counseling and psychotherapy. *Counseling and Values*, 33(3):164–176, 1989.
- [299] Monica Peddle, Lisa Mckenna, Margaret Bearman, and Debra Nestel. Development of non-technical skills through virtual patients for undergraduate nursing students: An exploratory study. *Nurse education today*, 73:94–101, 2019.
- [300] Mathijs Pellemans, Salim Salmi, Saskia Mérelle, Wilco Janssen, and Rob van der Mei. 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, 2024.
- [301] Shuang Peng, Shuai Zhu, Minghui Yang, Haozhou Huang, Dan Liu, Zujie Wen, Xuelian Li, and Biao Fan. Adacoach: A virtual coach for training customer service agents. *arXiv preprint arXiv:2204.12935*, 2022.
- [302] Joaquín Pérez, Eva Cerezo, Francisco J Serón, and Luis-Felipe Rodríguez. A cognitive-affective architecture for ecas. *Biologically Inspired Cognitive Architectures*, 18:33–40, 2016.
- [303] Leanne Picketts, Marika Dawn Warren, and Carrie Bohnert. Diversity and inclusion in simulation: addressing ethical and psychological safety concerns when working with simulated participants. *BMJ simulation & technology enhanced learning*, 7(6): 590, 2021.

- [304] Aaron Pico, Emilio Vivancos, Ana García-Fornes, and Vicent J Botti. Exploring text-generating large language models (llms) for emotion recognition in affective intelligent agents. In *ICAART (1)*, pages 491–498, 2024.
- [305] J David Pincus. Values as motives: Implications for theory, methods, and practice. *Integrative Psychological and Behavioral Science*, 58(4):1704–1750, 2024.
- [306] Jane Pirkis, Aves Middleton, Bridget Bassilios, Meredith Harris, Matthew J Spittal, Izabela Fedyszyn, Patty Chondros, and Jane Gunn. Frequent callers to telephone helplines: New evidence and a new service model. *International Journal of Mental Health Systems*, 10(1):43, 2016.
- [307] Isabella Poggi, Catherine Pelachaud, Fiorella de Rosis, Valeria Carofiglio, and Berardina De Carolis. Greta. a believable embodied conversational agent. *Multimodal intelligent information presentation*, pages 3–25, 2005.
- [308] Pooja Prajod, Mohammed Al Owayyed, Tim Rietveld, Jaap-Jan van der Steeg, and Joost Broekens. The effect of virtual agent warmth on human-agent negotiation. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 71–76, 2019.
- [309] Prolific. Researcher help center. <https://researcher-help.prolific.com/en/>, 2025. Accessed: 2026-01-13.
- [310] Slaviša Radović, Olga Firssova, Hans GK Hummel, and Marjan Vermeulen. Improving academic performance: Strengthening the relation between theory and practice through prompted reflection. *Active Learning in Higher Education*, 24(2):139–154, 2023.
- [311] AM Rahman, Abdullah Al Mamun, and Alma Islam. Programming challenges of chatbot: Current and future prospective. In *2017 IEEE region 10 humanitarian technology conference (R10-HTC)*, pages 75–78. IEEE, 2017.
- [312] Anand S Rao, Michael P Georgeff, et al. Bdi agents: From theory to practice. In *Icmas*, volume 95, pages 312–319, 1995.
- [313] Greg M Reger, Aaron M Norr, Patrick Sylvers, Jessica Peltan, Daniel Fischer, Matthew Trimmer, Shelan Porter, Pamela Gant, John S Baer, et al. 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, 2020.
- [314] Diego J Rivera-Gutierrez, Regis Kopper, Andrea Kleinsmith, Juan Cendan, Glen Finney, and Benjamin Lok. Exploring gender biases with virtual patients for high stakes interpersonal skills training. In *International Conference on Intelligent Virtual Agents*, pages 385–396. Springer, 2014.
- [315] Andrew Robb, Casey White, Andrew Cordar, Adam Wendling, Samsun Lampotang, and Benjamin Lok. A comparison of speaking up behavior during conflict with real and virtual humans. *Computers in Human Behavior*, 52:12–21, 2015.

- [316] Henry L Roediger and Andrew C Butler. The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences*, 15(1):20–27, 2011.
- [317] Carl Rogers. *Client centered therapy (new ed)*. Hachette UK, 2012.
- [318] Carl Ransom Rogers et al. *A theory of therapy, personality, and interpersonal relationships: As developed in the client-centered framework*, volume 3. McGraw-Hill New York, 1959.
- [319] Luke Rogers. Developing simulations in multi-user virtual environments to enhance healthcare education. *British Journal of Educational Technology*, 42(4):608–615, 2011.
- [320] Ido Roll, Deborah Butler, Nikki Yee, Ashley Welsh, Sarah Perez, Adriana Briseno, Katherine Perkins, and Doug Bonn. Understanding the impact of guiding inquiry: The relationship between directive support, student attributes, and transfer of knowledge, attitudes, and behaviours in inquiry learning. *Instructional Science*, 46(1):77–104, 2018.
- [321] Brent Rossen and Benjamin Lok. A crowdsourcing method to develop virtual human conversational agents. *International Journal of Human-Computer Studies*, 70(4): 301–319, 2012.
- [322] Geneviève Rouleau, Jérôme Pelletier, José Côté, Marie-Pierre Gagnon, Valérie Martel-Laferrrière, Rock Lévesque, SimforHealth, and Guillaume Fontaine. 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, 2020.
- [323] Amy Rubin, Nicholas A Livingston, Julianne Brady, Elise Hocking, Timothy Bickmore, Molly Sawdy, Nancy Kressin, Richard Saitz, and Steven Simon. 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, 2022.
- [324] Lilach Sagiv, Sonia Roccas, Jan Ciecich, and Shalom H Schwartz. Personal values in human life. *Nature human behaviour*, 1(9):630–639, 2017.
- [325] Michael Sailer and Lisa Homner. The gamification of learning: A meta-analysis. *Educational Psychology Review*, 32(1):77–112, 2020.
- [326] Salim Salmi, Saskia Mérelle, Renske Gilissen, and Willem-Paul Brinkman. Content-based recommender support system for counselors in a suicide prevention chat helpline: Design and evaluation study. *Journal of medical internet research*, 23(1): e21690, 2021.
- [327] Samiha Samrose and Ehsan Hoque. Mia: Motivational interviewing agent for improving conversational skills in remote group discussions. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–24, 2022.

- [328] Lindsay Sanneman and Julie A Shah. 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*, pages 94–110. Springer, 2020.
- [329] Evelyn-Rose Saus, Bjørn Helge Johnsen, Jarle Eid, Evelyn-Rose Saus, and Jarle Eid. Perceived learning outcome: The relationship between experience, realism and situation awareness during simulator training. *International maritime health*, 62(4): 258–264, 2010.
- [330] R Keith Sawyer. *The Cambridge handbook of the learning sciences*. Cambridge University Press, 2005.
- [331] Riley Scherr, Faris F Halaseh, Aidin Spina, Saman Andalib, and Ronald Rivera. Chatgpt interactive medical simulations for early clinical education: case study. *JMIR Medical Education*, 9:e49877, 2023.
- [332] Antoinette Schoenthaler, Glenn Albright, Judith Hibbard, Ron Goldman, et al. Simulated conversations with virtual humans to improve patient-provider communication and reduce unnecessary prescriptions for antibiotics: a repeated measure pilot study. *JMIR medical education*, 3(1):e6305, 2017.
- [333] Ricarda Scholz-Kuhn, Elena Makarova, Anat Bardi, Lukas F Litzellachner, Maya Benish-Weisman, and Anna K Döring. 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*, 40(3):75, 2025.
- [334] Dale H Schunk. *Learning theories an educational perspective*. Pearson Education, Inc, 2012.
- [335] S. H. Schwartz, B. Breyer, and D. Danner. Human values scale (ess). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*, 2015. doi: 10.6102/zis234. URL <https://zis.gesis.org/DoiId/zis234>.
- [336] Shalom H Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):11, 2012.
- [337] Neal E Seymour, Anthony G Gallagher, Sanziana A Roman, Michael K O’Brien, Vipin K Bansal, Dana K Andersen, and Richard M Satava. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Annals of surgery*, 236(4):458–464, 2002.
- [338] Hevil Shah, Brent Rossen, Benjamin Lok, Donna Londino, Scott D Lind, and Adriana Foster. Interactive virtual-patient scenarios: an evolving tool in psychiatric education. *Academic Psychiatry*, 36:146–150, 2012.
- [339] Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. Counseling-style reflection generation using generative pretrained transformers with augmented

- context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–20, 2020.
- [340] Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. *arXiv preprint arXiv:2311.09718*, 2023.
- [341] Valerie J Shute. Focus on formative feedback. *Review of educational research*, 78(1): 153–189, 2008.
- [342] Julika Siemer and Marios C Angelides. A comprehensive method for the evaluation of complete intelligent tutoring systems. *Decision support systems*, 22(1):85–102, 1998.
- [343] SIMmersion, N.d. URL <https://www.simmersion.com/>. Accessed 2023-09-11.
- [344] Trine Natasja Sindahl. *Chat Counselling for Children and Youth - A Handbook*. 2011.
- [345] Burrhus Frederic Skinner. *Science and human behavior*. Number 92904. Simon and Schuster, 1965.
- [346] SLO. Nederlands - schriftelijk onderwijs - kerndoel 5 - groep 5 en 6 - wat doen de kinderen. URL <https://www.slo.nl/thema/meer/tule/nederlands/kerndoel-5/?mode=groepen-56-leerling>.
- [347] Petr Slovák, Anja Thieme, David Murphy, Paul Tennent, Patrick Olivier, and Geraldine Fitzpatrick. On becoming a counsellor: Challenges and opportunities to support interpersonal skills training. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1336–1347, 2015.
- [348] Dragon Slumber. Tech support: Error unknown. [Steam], 2019.
- [349] Robbert Smit, Patricia Bachmann, Heidi Dober, and Kurt Hess. Feedback levels and their interaction with the mathematical reasoning process. *The Curriculum Journal*, 35(2):184–202, 2024.
- [350] Robert A Sottolare, Keith W Brawner, Anne M Sinatra, and Joan H Johnston. An updated concept for a generalized intelligent framework for tutoring (gift). *GIFTtutoring.org*, pages 1–19, 2017.
- [351] Rand J Spiro and Jihn-Chang Jehng. Cognitive flexibility and hypertext: Theory and technology for the nonlinear and multidimensional traversal of complex subject matter. In *Cognition, education, and multimedia*, pages 163–205. Routledge, 2012.
- [352] Herbert F Spitzer. Studies in retention. *Journal of Educational Psychology*, 30(9):641, 1939.

- [353] Francois St-Hilaire, Dung Do Vu, Antoine Frau, Nathan Burns, Farid Faraji, Joseph Potochny, Stephane Robert, Arnaud Roussel, Selene Zheng, Taylor Glazier, et al. A new era: Intelligent tutoring systems will transform online learning for millions. *arXiv preprint arXiv:2203.03724*, 2022.
- [354] Ian Steenstra, Farnaz Nouraei, Mehdi Arjmand, and Timothy Bickmore. Virtual agents for alcohol use counseling: Exploring llm-powered motivational interviewing. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*, pages 1–10, 2024.
- [355] Stichting De Kindertelefoon. De Kindertelefoon Jaarverslag 2023, 2023. URL <https://jaarverslag.kindertelefoon.nl/2023>. [Online; accessed 13-October-2024].
- [356] Stichting De Kindertelefoon. De Kindertelefoon Jaarverslag 2024, 2024. URL <https://jaarverslag.kindertelefoon.nl/2024>. [Online; accessed 13-October-2025].
- [357] Berglind Sveinbjörnsdóttir, Snorri Hjörvar Jóhannsson, Júlía Oddsdóttir, Tinna Þuríður Sigurðardóttir, Gunnar Ingi Valdimarsson, and Hannes Högni Vilhjálmsson. Virtual discrete trial training for teacher trainees. *Journal on Multimodal User Interfaces*, 13:31–40, 2019.
- [358] William Swartout, Benjamin D Nye, Arno Hartholt, Adam Reilly, Arthur C Graesser, Kurt VanLehn, Jon Wetzel, Matt Liewer, Fabrizio Morbini, Brent Morgan, et al. Designing a personal assistant for life-long learning (pal3). In *29th International Florida Artificial Intelligence Research Society Conference, FLAIRS 2016*, pages 491–496. AAAI press, 2016.
- [359] William R Swartout, Jonathan Gratch, Randall W Hill Jr, Eduard Hovy, Stacy Marsella, Jeff Rickel, and David Traum. Toward virtual humans. *AI Magazine*, 27(2):96–96, 2006.
- [360] John Sweller. Cognitive load theory. In *Psychology of learning and motivation*, volume 55, pages 37–76. Elsevier, 2011.
- [361] Nicolas Szilas, Lucie Chauveau, Kasper Andkjaer, Anna Laura Luiu, Mireille Bétran-court, and Frédéric Ehrler. Virtual patient interaction via communicative acts. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 91–93, 2019.
- [362] Niels Taatgen and John R Anderson. The past, present, and future of cognitive architectures. *Topics in Cognitive Science*, 2(4):693–704, 2010.
- [363] Hiroki Tanaka, Sakti Sakriani, Graham Neubig, Tomoki Toda, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. Teaching social communication skills through human-agent interaction. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 6(2):1–26, 2016.

- [364] Michael J Tanana, Christina S Soma, Vivek Srikumar, David C Atkins, and Zac E Imel. Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. *Journal of medical Internet research*, 21(7):e12529, 2019.
- [365] Kenneth W Thomas. Toward multi-dimensional values in teaching: The example of conflict behaviors. *Academy of management Review*, 2(3):484–490, 1977.
- [366] Nickolai Titov, Gavin Andrews, Isabella Choi, Genevieve Schwencke, and Alison Mahoney. Shyness 3: randomized controlled trial of guided versus unguided internet-based cbt for social phobia. *Australian & New Zealand Journal of Psychiatry*, 42(12):1030–1040, 2008.
- [367] Jennifer Tjia, Michele Pugnaire, Joanne Calista, Ethan Eisdorfer, Janet Hale, Jill Terrien, Olga Valdman, Stacy Potts, Maria Garcia, Majid Yazdani, et al. Using simulation-based learning with standardized patients (sp) in an implicit bias mitigation clinician training program. *Journal of medical education and curricular development*, 10:23821205231175033, 2023.
- [368] F Burcu Topu and Yuksel Goktas. The effects of guided-unguided learning in 3d virtual environment on students' engagement and achievement. *Computers in Human Behavior*, 92:1–10, 2019.
- [369] Matthew Turk. Multimodal interaction: A review. *Pattern recognition letters*, 36:189–195, 2014.
- [370] Udey. Prompt engineering for AI: Learn the art of prompting. <https://www.udemy.com/course/prompt-engineering-for-ai/>, 2024. Accessed: 2025-04-04.
- [371] Thomas Daniel Ullmann. Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education*, 29(2):217–257, 2019.
- [372] UNESCO. *AI and the Future of Education: Disruptions, Dilemmas and Directions*. UNESCO Publishing, Paris, France, 2025. URL <https://www.unesco.org/en/articles/ai-and-future-education-disruptions-dilemmas-and-directions>. Accessed: 2025-12-13.
- [373] Florina Uzefovsky, Anna K Döring, and Ariel Knafo-Noam. Values in middle childhood: Social and genetic contributions. *Social Development*, 25(3):482–502, 2016.
- [374] Frederik Vaassen, Jeroen Wauters, Frederik Van Broeckhoven, Maarten Van Overveldt, Walter Daelemans, and Koen Eneman. delearyou: Training interpersonal communication skills using unconstrained text input. *Proc. of ECGBL*, pages 505–513, 2012.
- [375] Fabienne M Van der Kleij, Remco CW Feskens, and Theo JHM Eggen. Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of educational research*, 85(4):475–511, 2015.

- [376] Rachel R Van Der Stuyf. Scaffolding as a teaching strategy. *Adolescent learning and development*, 52(3):5–18, 2002.
- [377] Jasper van der Waa, Jurriaan van Diggelen, Luciano Cavalcante Siebert, Mark Neerincx, and Catholijn Jonker. Allocation of moral decision-making in human-agent teams: A pattern approach. In *Engineering Psychology and Cognitive Ergonomics. Cognition and Design: 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*, pages 203–220. Springer, 2020.
- [378] Thomas L van der Weide. *Arguing to motivate decisions*. PhD thesis, Utrecht University, 2011.
- [379] JM Van der Zwaan, V Dignum, and CM Jonker. A bdi dialogue agent for social support: Specification and evaluation method. In *AAMAS 2012: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, Workshop on Emotional and Empathic Agents, Valencia, Spain, 4-8 June 2012; authors version*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2012.
- [380] Kurt VanLehn. The behavior of tutoring systems. *International journal of artificial intelligence in education*, 16(3):227–265, 2006.
- [381] Jacqueline Vaughn, Shannon H Ford, Melissa Scott, Carolyn Jones, and Allison Lewinski. Enhancing healthcare education: Leveraging chatgpt for innovative simulation scenarios. *Clinical Simulation in Nursing*, 87:101487, 2024.
- [382] Eva Veldhuizen-Ochodničánová, Róisín Cotton, Charlene Putney, Maggie Brennan, and Mads Haahr. The use of serious games in knowledge-and skills-based digital counselling: Applications to trafficking in persons in the war in ukraine. In *Joint International Conference on Serious Games*, pages 290–298. Springer, 2023.
- [383] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. User acceptance of information technology: Toward a unified view. *MIS quarterly*, pages 425–478, 2003.
- [384] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. The influence of interdependence and a transparent or explainable communication style on human-robot teamwork. *Frontiers in Robotics and AI*, 9:993997, 2022.
- [385] VirtualSpeech, N.d. URL <https://virtualspeech.com/>. Accessed 2023-09-11.
- [386] Matthew M Walsh, Michael A Krusmark, Tiffany Jastremski, Devon A Hansen, Kimberly A Honn, and Glenn Gunzelmann. Enhancing learning and retention through the distribution of practice repetitions across multiple sessions. *Memory & Cognition*, 51(2):455–472, 2023.
- [387] Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. Patient- ψ :

- Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*, 2024.
- [388] Micki Washburn, Danielle E Parrish, and Patrick S Bordnick. Virtual patient simulations for brief assessment of mental health disorders in integrated care settings. *Social Work in Mental Health*, 18(2):121–148, 2020.
- [389] John B Watson. *Behaviorism*. Routledge, 2017.
- [390] Lynelle Watts. Reflective practice, reflexivity, and critical reflection in social work education in australia. *Australian Social Work*, 72(1):8–20, 2019.
- [391] Ann Weatherall, Susan Danby, Karin Osvaldsson, Jakob Cromdal, and Michael Emmison. Pranking in children’s helpline calls. *Australian Journal of Linguistics*, 36(2):224–238, 2016.
- [392] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Posen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.
- [393] Cynthia Weston and Patricia A Cranton. Selecting instructional strategies. *The Journal of Higher Education*, 57(3):259–288, 1986.
- [394] Grant Wiggins. Seven keys to effective feedback. *Feedback*, 70(1):10–16, 2012.
- [395] Jason D Williams. The best of both worlds: Unifying conventional dialog systems and pomdps. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.
- [396] Benedikt Wisniewski, Klaus Zierer, and John Hattie. The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in psychology*, 10:487662, 2020.
- [397] Linjin Xi, Yi Zhang, and Qiyun Wang. Investigating the effects of an llm-based socratic conversational agent on students’ academic performance and reflective thinking in higher education. *Computers & Education*, page 105494, 2025.
- [398] Vakkas Yalçın. 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, 2021.
- [399] Chen Yang, Markus Gross, and Rafael Wampfler. Steering narrative agents through a dynamic cognitive framework for guided emergent storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 21, pages 377–387, 2025.
- [400] Heng Yao, Alexandre Gomes de Siqueira, Adriana Foster, Igor Galynker, and Benjamin Lok. Toward automated evaluation of empathetic responses in virtual human interaction systems for mental health scenarios. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2020.

- [401] Heng Yao, Alexandre Gomes de Siqueira, Anokhi Bafna, Devon Peterkin, Jenelle Richards, Megan L Rogers, Adriana Foster, Igor Galynker, and Benjamin Lok. A virtual human interaction using scaffolded ping-pong feedback for healthcare learners to practice empathy skills. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2022.
- [402] Heng Yao, Alexandre Gomes de Siqueira, Margeaux Johnson, Roberta Pileggi, Amy Blue, Michael D Bumbach, Rene Love, and Benjamin Lok. Enhancing empathic communication in healthcare education through virtual conversations: Leveraging large language models for real-time feedback. In *Proceedings of the 26th Symposium on Virtual and Augmented Reality*, pages 41–50, 2024.
- [403] Ugan Yasavur, Christine Lisetti, and Naphtali Rishe. Let’s talk! speaking virtual counselor offers you a brief intervention. *Journal on Multimodal User Interfaces*, 8(4): 381–398, 2014.
- [404] Hua Jonathan Ye and Atreyi Kankanhalli. Solvers’ participation in crowdsourcing platforms: Examining the impacts of trust, and benefit and cost factors. *The Journal of Strategic Information Systems*, 26(2):101–117, 2017.
- [405] Peijun Ye, Tao Wang, and Fei-Yue Wang. A survey of cognitive architectures in the past 20 years. *IEEE transactions on cybernetics*, 48(12):3280–3290, 2018.
- [406] Maryam Zahabi and Ashiq Mohammed Abdul Razak. Adaptive virtual reality-based training: a systematic literature review and framework. *Virtual Reality*, 24(4):725–752, 2020.
- [407] Eric D Zaizar, Michael A Gramlich, Albert Rizzo, Greg M Reger, Aaron M Norr, et al. 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*, 2024.
- [408] Sabrina Ziebarth, Anna Kizina, H Ulrich Hoppe, and Lorena Dini. A serious game for training patient-centered medical interviews. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*, pages 213–217. IEEE, 2014.
- [409] Marjorie A Zielke, Djakhangir Zakhidov, Daniel Jacob, and Gary Hardee. Beyond fun and games: toward an adaptive and emergent learning platform for pre-med students with the ut time portal. In *2016 IEEE International Conference on Serious Games and Applications for Health (SeGAH)*, pages 1–8. IEEE, 2016.
- [410] Marjorie A Zielke, Djakhangir Zakhidov, Daniel Jacob, and Gary Hardee. Beyond fun and games: toward an adaptive and emergent learning platform for pre-med students with the ut time portal. In *2016 IEEE International Conference on Serious Games and Applications for Health (SeGAH)*, pages 1–8. IEEE, 2016.
- [411] Leon Zlotos, Ailsa Power, Duncan Hill, and Paul Chapman. A scenario-based virtual patient program to support substance misuse education. *American journal of pharmaceutical education*, 80(3), 2016.

SIKS DISSERTATIONS

- 2016 01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
- 02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
- 03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
- 04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data
- 05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
- 06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
- 07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
- 08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
- 09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
- 10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
- 11 Anne Schuth (UvA), Search Engines that Learn from Their Users
- 12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
- 13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
- 14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
- 15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
- 16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
- 17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
- 18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
- 19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data
- 20 Daan Odijk (UvA), Context & Semantics in News & Web Search
- 21 Alejandro Moreno Céleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
- 22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems
- 23 Fei Cai (UvA), Query Auto Completion in Information Retrieval
- 24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
- 25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
- 26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
- 27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
- 28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
- 29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
- 30 Ruud Mattheij (TiU), The Eyes Have It
- 31 Mohammad Khelghati (UT), Deep web content monitoring
- 32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
- 33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
- 34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
- 35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation

-
- 36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
- 37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
- 38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
- 39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
- 40 Christian Detweiler (TUD), Accounting for Values in Design
- 41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
- 42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
- 43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
- 44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
- 45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
- 46 Jorge Gallego Perez (UT), Robots to Make you Happy
- 47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
- 48 Tanja Buttler (TUD), Collecting Lessons Learned
- 49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
- 50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains
-
- 2017 01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
- 02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
- 03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
- 04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
- 05 Mahdieh Shadi (UvA), Collaboration Behavior
- 06 Damir Vandić (EUR), Intelligent Information Systems for Web Product Search
- 07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
- 08 Rob Konijn (VUA), Detecting Interesting Differences: Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
- 09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
- 10 Robby van Delden (UT), (Steering) Interactive Play Behavior
- 11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
- 12 Sander Leemans (TU/e), Robust Process Mining with Guarantees
- 13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
- 14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
- 15 Peter Berck (RUN), Memory-Based Text Correction
- 16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines
- 17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
- 18 Ridho Reinanda (UvA), Entity Associations for Search
- 19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
- 20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
- 21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
- 22 Sara Magliacane (VUA), Logics for causal inference under uncertainty
- 23 David Graus (UvA), Entities of Interest – Discovery in Digital Traces
- 24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
- 25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
- 26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
- 27 Michiel Jooisse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors

- 28 John Klein (VUA), Architecture Practices for Complex Contexts
 - 29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT”
 - 30 Wilma Latuny (TiU), The Power of Facial Expressions
 - 31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
 - 32 Thaer Samar (RUN), Access to and Retrieval of Content in Web Archives
 - 33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
 - 34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
 - 35 Martine de Vos (VUA), Interpreting natural science spreadsheets
 - 36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
 - 37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
 - 38 Alex Kayal (TUD), Normative Social Applications
 - 39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
 - 40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
 - 41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
 - 42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
 - 43 Maaïke de Boer (RUN), Semantic Mapping in Video Retrieval
 - 44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
 - 45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
 - 46 Jan Schneider (OU), Sensor-based Learning Support
 - 47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
 - 48 Angel Suarez (OU), Collaborative inquiry-based learning
-
- 2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
 - 02 Felix Mannhardt (TU/e), Multi-perspective Process Mining
 - 03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
 - 04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
 - 05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
 - 06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
 - 07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
 - 08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
 - 09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
 - 10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
 - 11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
 - 12 Xixi Lu (TU/e), Using behavioral context in process mining
 - 13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
 - 14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters
 - 15 Naser Davarzani (UM), Biomarker discovery in heart failure
 - 16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
 - 17 Jianpeng Zhang (TU/e), On Graph Sample Clustering
 - 18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
 - 19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
 - 20 Manxia Liu (RUN), Time and Bayesian Networks
 - 21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games

- 22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
 - 23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
 - 24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
 - 25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
 - 26 Roelof Anne Jelle de Vries (UT), Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
 - 27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
 - 28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
 - 29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech
 - 30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web
-
- 2019 01 Rob van Eijk (UL), Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
 - 02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
 - 03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
 - 04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
 - 05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
 - 06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
 - 07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
 - 08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
 - 09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems
 - 10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
 - 11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
 - 12 Jacqueline Heinerman (VUA), Better Together
 - 13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
 - 14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
 - 15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
 - 16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
 - 17 Ali Hurriyetoglu (RUN), Extracting actionable information from microtexts
 - 18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
 - 19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
 - 20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
 - 21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
 - 22 Martin van den Berg (VUA), Improving IT Decisions with Enterprise Architecture
 - 23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
 - 24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
 - 25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
 - 26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
 - 27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
 - 28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
 - 29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
 - 30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
 - 31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
 - 32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
 - 33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks

-
- 34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
- 35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming
- 36 Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
- 37 Jian Fang (TUD), Database Acceleration on FPGAs
- 38 Akos Kadar (OU), Learning visually grounded and multilingual representations
-
- 2020 01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
- 02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
- 03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
- 04 Maarten van Gompel (RUN), Context as Linguistic Bridges
- 05 Yulong Pei (TU/e), On local and global structure mining
- 06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
- 07 Wim van der Vegt (OU), Towards a software architecture for reusable game components
- 08 Ali Mirsoleimani (UL), Structured Parallel Programming for Monte Carlo Tree Search
- 09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
- 10 Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
- 11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation Methods for Long-Tail Entity Recognition Models
- 12 Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
- 13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
- 14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
- 15 Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
- 16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
- 17 Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
- 18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
- 19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
- 20 Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
- 21 Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
- 22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
- 23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
- 24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
- 25 Xin Du (TU/e), The Uncertainty in Exceptional Model Mining
- 26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer optimization
- 27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
- 28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
- 29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
- 30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
- 31 Gongjin Lan (VUA), Learning better - From Baby to Better
- 32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
- 33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
- 34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development

-
- 35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production
-
- 2021 01 Francisco Xavier Dos Santos Fonseca (TUD), Location-based Games for Social Interaction in Public Space
- 02 Rijk Mercurur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
- 03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
- 04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
- 05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
- 06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
- 07 Armel Lefebvre (UU), Research data management for open science
- 08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
- 09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
- 10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
- 11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
- 12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
- 13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
- 14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
- 15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
- 16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
- 17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
- 18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
- 19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
- 20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
- 21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
- 22 Sihang Qiu (TUD), Conversational Crowdsourcing
- 23 Hugo Manuel Proença (UL), Robust rules for prediction and description
- 24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
- 25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
- 26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
- 27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
- 28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs
-
- 2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
- 02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
- 03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
- 04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
- 05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
- 06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
- 07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
- 08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
- 09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
- 10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines

- 11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
 - 12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
 - 13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
 - 14 Michiel Overeem (UU), Evolution of Low-Code Platforms
 - 15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
 - 16 Pieter Gijbbers (TU/e), Systems for AutoML Research
 - 17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
 - 18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
 - 19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
 - 20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
 - 21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
 - 22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
 - 23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
 - 24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
 - 25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
 - 26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
 - 27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
 - 28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
 - 29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
 - 30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
 - 31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
 - 32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
 - 33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
 - 34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
 - 35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction
-
- 2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
 - 02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
 - 03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
 - 04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
 - 05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications
 - 06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
 - 07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
 - 08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
 - 09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
 - 10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
 - 11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
 - 12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
 - 13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
 - 14 Selma Čaušević (TUD), Energy resilience through self-organization
 - 15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models

-
- 16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
- 17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
- 18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
- 19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
- 20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
- 21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
- 22 Alireza Shojaifar (UU), Volitional Cybersecurity
- 23 Theo Theunissen (UU), Documentation in Continuous Software Development
- 24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
- 25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
- 26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
- 27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
- 28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
- 29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results
-
- 2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
- 02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems
- 03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
- 04 Mike Huisman (UL), Understanding Deep Meta-Learning
- 05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
- 06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence
- 07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
- 08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
- 09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
- 10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
- 11 withdrawn
- 12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
- 13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
- 14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
- 15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
- 16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
- 17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
- 18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
- 19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
- 20 Ritsart Anne Plantenga (UL), Omgang met Regels
- 21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
- 22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
- 23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
- 24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
- 25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
- 26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework

- 27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning
 - 28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs
 - 29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
 - 30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
 - 31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
 - 32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
 - 33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction
 - 34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method
 - 35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach
 - 36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
 - 37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
 - 38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings
 - 39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
 - 40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
 - 41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines
 - 42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
 - 43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms
 - 44 Marcel Schmitz (OU), Towards learning analytics-supported learning design
 - 45 Sara Salimzadeh (TUD), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making
 - 46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology
 - 47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs
 - 48 Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions
-
- 2025 01 Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems
 - 02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks
 - 03 Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI
 - 04 Rianne Schouten (TU/e), Exceptional Model Mining for Hierarchical Data
 - 05 Nele Albers (TUD), Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation
 - 06 Daniël Vos (TUD), Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization
 - 07 Ricky Maulana Fajri (TU/e), Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance
 - 08 Stefan Bloemheuvel (TiU), Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction
 - 09 Fadime Kaya (VUA), Decentralized Governance Design - A Model-Based Approach
 - 10 Zhao Yang (UL), Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning
 - 11 Shahin Sharifi Noorian (TUD), From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration
 - 12 Lijun Lyu (TUD), Interpretability in Neural Information Retrieval
 - 13 Fuda van Diggelen (VUA), Robots Need Some Education: on the complexity of learning in evolutionary robotics
 - 14 Gennaro Gala (TU/e), Probabilistic Generative Modeling with Latent Variable Hierarchies
 - 15 Michiel van der Meer (UL), Opinion Diversity through Hybrid Intelligence
 - 16 Monika Grewal (TU Delft), Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging
 - 17 Matteo De Carlo (VUA), Real Robot Reproduction: Towards Evolving Robotic Ecosystems

- 18 Anouk Neerinx (UU), Robots That Care: How Social Robots Can Boost Children's Mental Wellbeing
- 19 Fang Hou (UU), Trust in Software Ecosystems
- 20 Alexander Melchior (UU), Modelling for Policy is More Than Policy Modelling (The Useful Application of Agent-Based Modelling in Complex Policy Processes)
- 21 Mandani Ntekouli (UM), Bridging Individual and Group Perspectives in Psychopathology: Computational Modeling Approaches using Ecological Momentary Assessment Data
- 22 Hilde Weerts (TU/e), Decoding Algorithmic Fairness: Towards Interdisciplinary Understanding of Fairness and Discrimination in Algorithmic Decision-Making
- 23 Roderick van der Weerd (VUA), IoT Measurement Knowledge Graphs: Constructing, Working and Learning with IoT Measurement Data as a Knowledge Graph
- 24 Zhong Li (UL), Trustworthy Anomaly Detection for Smart Manufacturing
- 25 Kyana van Eijndhoven (TiU), A Breakdown of Breakdowns: Multi-Level Team Coordination Dynamics under Stressful Conditions
- 26 Tom Pepels (UM), Monte-Carlo Tree Search is Work in Progress
- 27 Danil Provodin (JADS, TU/e), Sequential Decision Making Under Complex Feedback
- 28 Jinke He (TU Delft), Exploring Learned Abstract Models for Efficient Planning and Learning
- 29 Erik van Haeringen (VUA), Mixed Feelings: Simulating Emotion Contagion in Groups
- 30 Myrthe Reuver (VUA), A Puzzle of Perspectives: Interdisciplinary Language Technology for Responsible News Recommendation
- 31 Gebrekirstos Gebreselassie Gebremeskel (RUN), Spotlight on Recommender Systems: Contributions to Selected Components in the Recommendation Pipeline
- 32 Ryan Brate (UU), Words Matter: A Computational Toolkit for Charged Terms
- 33 Merle Reimann (VUA), Speaking the Same Language: Spoken Capability Communication in Human-Agent and Human-Robot Interaction
- 34 Eduard C. Groen (UU), Crowd-Based Requirements Engineering
- 35 Urja Khurana (VUA), From Concept To Impact: Toward More Robust Language Model Deployment
- 36 Anna Maria Wegmann (UU), Say the Same but Differently: Computational Approaches to Stylistic Variation and Paraphrasing
- 37 Chris Kamphuis (RUN), Exploring Relations and Graphs for Information Retrieval
- 38 Valentina Maccatrozzo (VUA), Break the Bubble: Semantic Patterns for Serendipity
- 39 Dimitrios Alivanistos (VUA), Knowledge Graphs & Transformers for Hypothesis Generation: Accelerating Scientific Discovery in the Era of Artificial Intelligence
- 40 Stefan Grafberger (UvA), Declarative Machine Learning Pipeline Management via Logical Query Plans
- 41 Mozghan Vazifehdostirani (TU/e), Leveraging Process Flexibility to Improve Process Outcome - From Descriptive Analytics to Actionable Insights
- 42 Margherita Martorana (VUA), Semantic Interpretation of Dataless Tables: a metadata-driven approach for findable, accessible, interoperable and reusable restricted access data
- 43 Krist Shingjergji (OU), Sense the Classroom - Using AI to Detect and Respond to Learning-Centered Affective States in Online Education
- 44 Robbert Reijnen (TU/e), Dynamic Algorithm Configuration for Machine Scheduling Using Deep Reinforcement Learning
- 45 Anjana Mohandas Sheeladevi (VUA), Occupant-Centric Energy Management: Balancing Privacy, Well-being and Sustainability in Smart Buildings
- 46 Ya Song (TU/e), Graph Neural Networks for Modeling Temporal and Spatial Dimensions in Industrial Decision-making
- 47 Tom Kouwenhoven (UL), Collaborative Meaning-Making. The Emergence of Novel Languages in Humans, Machines, and Human-Machine Interactions
- 48 Evy van Weelden (TiU), Integrating Virtual Reality and Neurophysiology in Flight Training
- 49 Selene Báez Santamaría (VUA), Knowledge-centered conversational agents with a drive to learn
- 50 Lea Krause (VUA), Contextualising Conversational AI
- 51 Jiaxu Zhao (TU/e), Understanding and Mitigating Unwanted Biases in Generative Language Models
- 52 Qiao Xiao (TU/e), Model, Data and Communication Sparsity for Efficient Training of Neural Networks
- 53 Gaole He (TUD), Towards Effective Human-AI Collaboration: Promoting Appropriate Reliance on AI Systems

-
- 54 Go Sugimoto (VUA), MISSING LINKS Investigating the Quality of Linked Data and its Tools in Cultural Heritage and Digital Humanities
- 55 Sietze Kai Kuilman (TUD), AI that Glitters is Not Gold: Requirements for Meaningful Control of AI Systems
- 56 Wijnand van Woerkom (UU), A Fortiori Case-Based Reasoning: Formal Studies with Applications in Artificial Intelligence and Law
- 57 Syeda Amna Sohail (UT), Privacy-Utility Trade-Off in Healthcare Metadata Sharing and Beyond: A Normative and Empirical Evaluation at Inter and Intra Organizational Levels
- 58 Junhan Wen (TUD), "From iMage to Market": Machine-Learning-Empowered Fruit Supply
- 59 Mohsen Abbaspour Onari (TU/e), From Explanation to Trust: Modeling and Measuring Trust in Explainable Decision Support
- 60 Marcel Jurriaan Robeer (UU), Beyond Trust: A Causal Approach to Explainable AI in Law Enforcement
- 61 Shuai Wang (VUA), Links in Large Integrated Knowledge Graphs: Analysis, Refinement, and Domain Applications
- 62 Khaleel Asyraf Mat Sanusi (OU), Augmenting a learning model within immersive learning environments for psychomotor skills
- 63 Rashid Zaman (TU/e), Online Conformance Checking on Degraded Data
- 64 Jens d'Hondt (TU/e), Effective and Efficient Multivariate Similarity Search
- 65 Aswin Balasubramaniam (UT), Disentangling Runner Drone Interaction Potentialities
-
- 2026 01 Pei-Yu Chen (TUD), Human-Agent Alignment Dialogues: Eliciting User Information at Runtime for Personalized Behavior Support
- 02 Hezha Hassan Mohammedkhan (TiU), Estimating Body Measurements of Children from 2D Images: Towards the Automatic Detection of Malnutrition
- 03 Kyriakos Psarakis (TUD), Democratizing Scalable Cloud Applications: Transactional Stateful Functions on Streaming Dataflows
- 04 Boyu Xu (UU), Exploring Indirect Relations Between Topics in Neuroscience Literature Using Augmented Reality to Inform Experimental Design
- 05 Koen Minartz (TU/e), Stochastic Simulation with Geometric Deep Generative Models
- 06 Azim Afroozeh (CWI, VUA), FastLanes: A Next-Gen File Format
- 07 Inès Blin (VUA), Narrative Understanding with Knowledge Graphs
- 08 Paul van Vulpen (UU), Debating Digital Dominance: Decentralized Technology Governance For Strategic Autonomy
- 09 Afrizal Doewes (TU/e), Rethinking Automated Essay Scoring: Agreement, Fairness, and Feedback
- 10 Nikolaos Delapaschos Kondylidis (VUA), Establishing Task-Oriented Understanding between Agents
- 11 Işıl Baysal Erez (UT), Handling Missing Data with Meta-Learning and Large Language Models
- 12 Xue Li (UvA), From Fine-tuning to Prompting: A Paradigm Shift in Knowledge Graph Construction
- 13 Isaac da Silva Torres (VUA), Guidelines To Flux Between Conceptual Models: Understanding Complex Digital Business Ecosystems
- 14 Philip Lippmann (TUD), Synthetic Data for Robust Language Modelling
- 15 Rashmi Khazanchi (OU), Artificial Intelligence in Education: Impact of AI-Based Systems on Mathematics Achievement
- 16 Carolina Ferreira Gomes Centeio Jorge (TUD), Modelling Artificial Trust for Effective Human-AI Teamwork
- 17 Maria Tsfasman (TUD), Towards Predicting Memory in Multimodal Group Interactions
- 18 Riccardo Lo Bianco (TU/e), Deep Reinforcement Learning for Automated Decision-Making in Process Management Systems
- 19 Israel Campero Jurado (TU/e), Innovations in Optimization and Applications in Healthcare
- 20 Ifitahu Ni'mah (TU/e), Contrastive Learning and Evaluation in Low Resource Scenario of Natural Language Processing
- 21 Francisco N.F.Q. Simoes (UU), Causality, Information, and Decision-Making
- 22 Ruben Verhagen (TUD), Transparent and Explainable Agents for Human-Agent Teaming
- 23 Eduardo Calò (UU), Automatically Expressing the Meaning of Logical Formulae in Natural Language
- 24 Shuai Han (UU), Improving Sample Efficiency of Reinforcement Learning: Exploiting Structural Knowledge for Decision Making

- 25 Francis Saa-Dittoh (VUA), From Radio to AI: African Community-Driven Development of Sustainable Information Systems
- 26 Mohammed Al Owayyed (TUD), Interactive Simulation-Based Learning Tools for Training Children's Helpline Counsellors

ACKNOWLEDGMENTS

The story behind my PhD journey, I like to think, started in primary school. A teacher once asked us what we wanted to be in the future. To me, being a teacher sounded appealing; I liked explaining things and delving (I wrote “delving” without ChatGPT’s help) deeper into topics I’m interested in. The teacher laughed and suggested I choose something that pays more, since he thought I was a “smart student with a bright future.” When asked for an alternative, my brilliant reply was: a firefighter! I think that’s when he lost hope. Fast forward to my BSc days, when friends and colleagues expected me to pursue a PhD. I fit the stereotype: glasses, sitting in the first row, answering everyone’s questions, and (apparently) explaining complex things clearly. At home, my parents supported this PhD path, as both are educators (one with a PhD, the other working in a school). So I’m not sure whether expectations shaped me into doing a PhD, whether I shaped everyone’s expectations, or whether I was just avoiding the job market and staying in my comfort zone by continuing my graduate studies. Anyway, I’ll stop rambling. I’m proud, happy, and a bit sad to reach the end of this journey (and the start of the next). Many people made these four years easier, more enjoyable, and unforgettable—so I acknowledge them here, so they feel acknowledged, and I can officially be an acknowledger of the acknowledgedes.

Here, I want to start with my supervisors, Myrthe and Willem-Paul, who supported me academically, professionally, and personally. It felt like I was the main character in a story, while they served as guides, providing hints and structure along the way. Myrthe, we first met at my MSc thesis defence, and you were kind then and remain so. I admire your organisation and your care for well-being alongside research progress. I remember you encouraging me to take time off when I hadn’t in a while—I wasn’t convinced at the time, but it was absolutely the right call. Thank you for your support, creativity, and the (slightly challenging) discussions. Willem-Paul, you supervised my MSc thesis before my PhD, and you agreed to take me back for a PhD. Thank you for always going the extra mile, sometimes letting meetings run twice as long, and for your enthusiastic humour and memorable examples. Your ability to quickly pinpoint misunderstandings and structure clarifications helped me grow a lot as a researcher and as a person.

Additionally, I would like to thank my committee members: Maaïke Kleinsmann, Tibor Bosse, Renske Gilissen, Timothy Bickmore and Mark Neerinx. Thank you for your time and insights, and for attending the defence.

Next, I would like to thank the members of our group. I’m lucky to be part of Interactive Intelligence—everyone was genuinely nice. I made colleagues from officemates, friends from colleagues, and great memories from what is (kind of) a job. First, my paranymphs: Nele and Micha. Nele, although we have almost no shared interests (apart from research), we always find something new to talk about—I will not forget talking about video games while you had a confused look on your face. Thanks for your continuous help and for being a great “distraction collaborator” (your words). Micha, a five-minute conversation with

you can cover three topics, usually including politics, and it always ends with laughter, one way or another. I'm honoured that you generously added me to your phone contacts.

Kicking things off, with the members of the great office 4.W.920: Enrico, thanks for supporting me and pushing me outside my comfort zone at conferences and meetings (and, of course, for the Pokémon and Layton discussions); Shubhalaxmi, for the small talks and cooking nights; Zuzanna, for being up to talk about anything (and requesting forms); Urja, for surviving full office days; Mani (honorary member), for knowing everything, your witty humour, and your contributions to the office board; Nele (see above); and, of course, the Super Duck. Moving now to the office I was (kind of) meant to be part of: Carolina, thanks for the energy you brought to the group and our social gatherings, and for being great. Ruben, I'm impressed by your knowledge of the Saudi league, and I really enjoyed our Nintendo- and football-related talks. Sid, thanks for your guidance and kindness (and the delicious Golgappa). Pei-Yu, for being a reliable double-coder and a great colleague. Linyun, thanks for your infectious office energy, your corner-jump scares, and for always being up for any social activity (even Mario Kart). Continuing with others in the II group, ordered by proximity to office 4.W.920: Masha, for the fun talks about Ghibli, Japan, Nintendo, and life in general; Micha (see above); Amir, for the tennis games, and the football discussions; Sietze, for the video game discussions and for making me double-question my research value when I started; Antonio and Koray, for the friendly conversations; Paul, for the many video game-related discussions and for kayaking while I rested; Morita, for asking important questions and for your continuous support with double coding and discussions; Ruben (new Ruben), for always dropping by our office; Agnes and Charlotte, for adding energy and fun to the social/game nights; and Deborah and Shambhawi, for the fun conversations. I also want to thank the staff members of our group for their kindness and friendly chats: Catholijn, Mark, Pradeep, Catha, Luciano, and Stephanie. Special thanks to Anita for keeping everything running, Ruud for quick technical help, and Wouter for the nice chats.

Yet, one highlight of my PhD is the weekly meetings with postdocs, PhDs, and MSc students—it wouldn't have been the same without you. So thanks to Esra for your energy and for being a responsible and reliable double coder; Sarah, for helping me think from different perspectives in my experiments; Jennie—we still didn't get the promised cake; Alexander, for the interesting discussions about education with LLMs; and Amal, for your helpful feedback on my experiment setup. Also, I won't forget the (previously mentioned) Mani, Micha, and Nele (for the third time).

Over the course of my PhD, I supervised many talented MSc students who, hopefully, benefited, but who also taught me valuable lessons about supervision, time management, and understanding training systems in practice (the names are ordered chronologically, not by favourites). Ayrton and Dongxu, you examined Lilobot with direct feedback and an emotional model. At that time, in my first year, I was learning with you what difficulties to expect from updating Lilobot. It was a pleasure supervising you, and I enjoyed talking about other topics like video games and anime (I also enjoyed discussing the thesis, of course). Miriam, although I only supervised you in your final stage, the academic contribution was all yours, and I hope my continuous pushing to wrap up helped. Maya and Adarsh, you updated Lilobot with an overall session structure and added LLMs to address some of the limitations of the rule-based system. I enjoyed supervising you and our conversations. I

would also like to thank the BSc software project, honour project students, and student assistants, who, through their contributions and valuable skills, not only achieved great results but also improved and implemented parts of Lilobot, making it a stable environment for further development and future work. So thanks to Pablo, Alex, Fanni, Sten, Francisco, Irene, Kayleigh, Fedde, Veena, Tristan, Ashma, Vlad, Nikola, Nuno, and Meenakshi. I hope you benefited from this experience as much as I did—and if not, you can at least be proud that you contributed to a training system that will become part of an online course =)

Ultimately, I didn't think I would write acknowledgements this long (three pages so far?). Anyway, there are still some people I would like to thank. First, those who helped before I even started: Fran, for supervising my MSc thesis; and Sharon, for starting the Lilobot simulation during your MSc thesis—I hope you are proud of how much the project has grown. I would also like to thank De Kindertelefoon for their continuous support, and the ProtectMe consortium for bringing together enthusiastic people and inspiring discussions from different fields and perspectives. I would also like to thank my co-authors, who helped shape my work (or, hopefully, let me help shape theirs): Loes, Marcus, Arno, Kathleen, Chadha (or Shatha? at this point, I'm too embarrassed to ask), Merijn, and Niko. I would also like to thank the Saudi Cultural Mission for their continuous support and quick responses whenever I needed help, both personally and academically. Also, our neighbour and family friend, Kees, who passed away recently, I will never forget the best apple pie I have ever had. Speaking of food, there are other things that kept me going during my PhD: my favourite restaurants in Delft (Lychee, San Marco, Betty Boop, and Mekhong Thai), Nintendo and their amazing games, Yorkshire tea (with mint leaves), the “XD” emoji... and I think I should stop here...

!. Dear reader—if you have made it this far, there is a hidden message for you somewhere above. I'll let you figure it out. Hint: there is a reason this paragraph starts with an exclamation mark.

Finally, to friends and family back home: for friends, there are three groups, totalling 22 names, so I will spare the reader the full list (but you know yourselves). I would also like to thank my five siblings for their continuous support and several sibling fights: Saad (academic support), Owayyed (financial support), Mosa (logistical support), Sarah (with an h; PR support), and Yara (entertainment support?). To my parents as well: your continuous, unwavering support for my learning and my life has made me who I am today. Thank you for always standing by my decisions and offering all kinds of help and support, even without me asking. I hope I made you proud, as much as I am proud of you. Finally, to my wife and the love of my life, Hanan. You were the greatest support throughout this journey (and beyond). When I was struggling (only a bit) with my master's, we started living together; my performance and grades then became excellent, and they have remained so ever since! I know you think you need to repay me after I helped you with your MSc research, but I still think I haven't paid my dues and that I am the one in debt. Thank you, and I love you.

CURRICULUM VITÆ

Mohammed AL OWAYYED

EDUCATION

- 2022–2026 **Ph.D. in Computer Science**
Delft University of Technology, Delft, the Netherlands
Supervisors: Prof. Willem-Paul Brinkman and Dr. Myrthe Tielman
- 2018–2020 **Master of Science in Computer Science**
Delft University of Technology, Delft, the Netherlands
- 2012–2016 **Bachelor of Science in Software Engineering**
King Saud University, Riyadh, Saudi Arabia

EXPERIENCE

- 2022–2026 **Delft University of Technology**, Delft, the Netherlands
Ph.D. Candidate (Interactive Intelligence Group)
- 2016–Present **King Saud University**, Riyadh, Saudi Arabia
Lecturer and TA (Department of Software Engineering)
- 2015 **IBM IoT Lab**, Durham, NC, USA
Intern

LIST OF PUBLICATIONS

UNDER REVIEW

1. **Al Owayyed, M.**, Tielman, M. L., & Brinkman, W.P. (2025). *Simulation-based training for child helpline counsellors: Impact of simulation, feedback, and reflection in a randomised controlled trial*. Manuscript is under review.

2026

- 1 **Al Owayyed, M.**, Brinkman, W.P., Guan, K. W., 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>

2025

1. **Al Owayyed, M.**, Denga, A. A., & Brinkman, W.P. (2025). *Controlled yet natural: A hybrid BDI-LLM conversational agent for child helpline training*. In *IVA '25: Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents* (Article 17). ACM. <https://doi.org/10.1145/3717511.3747075>
2. 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), Article 5. <https://doi.org/10.1007/s10916-024-02121-8>
3. **Al Owayyed, M.**, Tielman, M. L., & Brinkman, W.P. (2025). *A simulation-based training tool for child helpline counsellors* [Conference demo abstract]. In *Proceedings of the 14th Supporting Health by Technology Conference*, Enschede, The Netherlands, p. 93.

2024

1. **Al Owayyed, M.**, Despan, A., Tielman, M. L., & Brinkman, W.P. (2024). *A cognitive conversational agent for training child helpline volunteers* [Conference demo abstract]. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents (IVA '24)* (Article 45). ACM. <https://doi.org/10.1145/3652988.3696197>
2. **Al Owayyed, M.**, Tielman, M. L., Hartholt, A., Specht, M. M., & Brinkman, W.P. (2024). *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>
3. **Al Owayyed, M.**, Tielman, M. L., Hartholt, A., Specht, M. M., & Brinkman, W.P. (2024). *Agent-based social skills training systems: The ARTES architecture* [Encore abstract]. In *Proceedings of the Joint International Scientific Conferences on AI and Machine Learning (BNAIC/BeNeLearn 2024)*, Utrecht, The Netherlands.

2023

1. **Al Owayyed, M.**, Grundmann, S., Bruijnes, M., & Brinkman, W.P. (2023). *Training child helpline counsellors with a BDI-based conversational agent* [Conference demo abstract]. In *Proceedings of the Joint International Scientific Conferences on AI and Machine Learning (BNAIC/BeNe-Learn 2023)*, Delft, The Netherlands.
2. Degachi, C., **Al Owayyed, M.**, & Tielman, M. L. (2023). *Trust and perceived control in burnout support chatbots*. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI 2023)*. ACM. <https://doi.org/10.1145/3544549.3585780>

2022


1. **Al Owayyed, M.**, Tielman, M. L., & Brinkman, W.P. (2022). *Virtual patients to train communication skills of healthcare providers* [Doctoral Consortium]. In *Proceedings of the 22nd Intelligent Virtual Agents Conference*, Faro, Portugal.

2020

1. **Al Owayyed, M. A. S.** (2020). *Motivating PhD candidates with depression symptoms to complete thoughts-strengthening exercises via a conversational agent* [Master's thesis, Delft University of Technology]. <https://resolver.tudelft.nl/uuid:c5ed5bde-e118-4b73-9f4b-85531920f19a>

2019

1. Prajod, P., **Al Owayyed, M.**, Rietveld, T., van der Steeg, J.-J., & Broekens, J. (2019). *The effect of virtual agent warmth on human-agent negotiation*. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*. IFAAMAS.
2. Mac An Bhaird, L., **Al Owayyed, M.**, van Driel, R., Jiang, H., Johannessen, R. A., Salamon, N. Z., Balint, J. T., & Bidarra, R. (2019). *Learning geothermal energy basics with the serious game HotPipe*. In *Proceedings of the Games and Learning Alliance Conference*.

 Included in this Thesis.

