Memory with Meaning
Enabling Value-Centric Long-Term Human-Agent Dialogue

Saveur, Tom; Axelsson, Agnes; Burger, Franziska; Neerincx, Mark; Oertel, Catharine

DOI
10.1145/3652988.3673925

Publication date
2024

Document Version
Final published version

Published in
IVA '24

Citation (APA)
Saveur, T., Axelsson, A., Burger, F., Neerincx, M., & Oertel, C. (2024). Memory with Meaning: Enabling Value-Centric Long-Term Human-Agent Dialogue. In R. Jack, M. Chollet, R. Aylett, T. Bickmore, S. Marsella, & G. Lucas (Eds.), *IVA '24: Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents* Article 37 ACM. https://doi.org/10.1145/3652988.3673925

Memory with Meaning: Enabling Value-Centric Long-Term Human-Agent Dialogue

Tom Saveur
T.Saveur@student.tudelft.nl
Delft University of Technology
Delft, South Holland, Netherlands

Agnes Axelsson
Franziska Burger
A.Axelsson@tudelft.nl
F.V.Burger@tudelft.nl
Delft University of Technology
Delft, South Holland, Netherlands

Mark Neerincx
Catharine Oertel
M.A.Neerincx@tudelft.nl
C.R.M.M.Oertel@tudelft.nl
Delft University of Technology
Delft, South Holland, Netherlands

ABSTRACT

When a human makes a decision, an observer may want to understand the reasons and motivations behind the decision. This understanding is important when IVAs are involved in contextual decision-making or coaching practices. To address this challenge, we propose that an agent's understanding of its user should include knowledge of the user's underlying *values*. Humans prioritise different values – sometimes contradictory – in a manner that depends on the context. We present a method where the agent and user build the required context-sensitive value model *together*. We use Schwartz's value theory, which places individuals' values into ten categories. In a between-subject experiment, with three sessions on different days, we *elicit* user values by presenting them with moral dilemmas in different contexts on the first day, *refine* the model by asking users to argue about contradictions on the second day, and let them *reflect* on the model that they have built together with the system on the third day. We find that users exposed to a value-aware condition are more likely to agree with the robot's representations of their values post-reflection than those in a baseline. Participants also prioritise different values depending on the context, agreeing with previous findings.

CCS CONCEPTS

• **Human-centered computing** → **User models**; *User studies*; *Natural language interfaces*; • **Applied computing** → Psychology.

KEYWORDS

Schwartz, values, human-robot interaction, value-aware systems, contextualised values, context, school, home, social, conformity, achievement, hedonism, self-direction, benevolence

ACM Reference Format:
Tom Saveur, Agnes Axelsson, Franziska Burger, Mark Neerincx, and Catharine Oertel. 2024. Memory with Meaning: Enabling Value-Centric Long-Term Human-Agent Dialogue. In *ACM International Conference on Intelligent Virtual Agents (IVA '24), September 16–19, 2024, GLASGOW, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3652988.3673925

IVA '24, September 16–19, 2024, GLASGOW, United Kingdom

ACM ISBN 979-8-4007-0625-7/24/09
https://doi.org/10.1145/3652988.3673925

1 INTRODUCTION

When an IVA gives decision support to a human, what does it need to know about the individual it is helping to give a good answer? In simpler decisions, like advice for cooking, there may be an objectively correct answer, and the system does not need to know anything about the user to give good advice. On the other hand, in **complex domains and contexts** – like disaster rescue or triage – user choices and actions can only be understood by mapping the users' actions to their belief system. If we are to build IVAs that assist in such complex cases, then the agent's responses and collaborative actions will improve from being contextual to the users and their abilities and values. A generic answer or action will not suffice.

We see a research gap in how IVAs can build an understanding of their user to support complex decisions. In order to address it, we give our agent **value-driven memory** to let it build an understanding of what the user believes and why. Values are a general way to represent the deeply held beliefs of humans in their most conceptual form [29, 31]. While values are supposed to be universally applicable, previous research also indicates that the selection of values that are relevant to an individual at a given point in time (*activated*) is context-dependent [19, 23, 38]. This would mean that the context affects the implementation of the value-driven memory. We therefore investigate whether participants indeed prioritise different values depending on context.

In this paper, we investigate whether building the robot's memory around participants' values is (i) noticeable to participants and (ii) if doing so leads to more value-aligned conversations. We further (iii) investigate whether participants prioritise values differently depending on context.

2 RELATED WORK

2.1 Value models

Schwartz proposed a set of universal values that motivate humans' "ideologies, attitudes, and actions in the political, religious, environmental and other domains" [29]. The evaluation by Schwartz confirmed that individuals from 20 countries had values that could be categorised as based on **benevolence**, **tradition**, **conformity**, **security**, **power**, **achievement**, **hedonism**, **stimulation**, **self-direction**, as well as **universalism**. When the 10 values are distributed in this order around a circle, values that are related to each other neighbour each other. For example, *security* and *power* are adjacent in Schwartz's circle because some participants expressed that they highly valued the ability to control uncertain situations.

Similar arguments are found for other pairs [29]. Schwartz et al. also propose grouping the circle into quadrants [29] and halves [31] respectively, representing higher-level reasons why a person may hold values. There is small body of work that proposes methods for how an individual's values can inform arguments for behaviour change, although this is not then used to create a memory model containing the values [8, 9].

2.1.1 Context-dependent values. Schwartz [30] calls the values we inherit from social context and society our *mental programs* [19, cf.]. Each individual may carry multiple mental programs, representing different and possibly contradictory sources of values like religion, age group, gender, political beliefs and one's job. When an individual makes a decision in a group of their friends, certain values are more easily activated than others in their possibly conflicting mental programs, and the person may make a different decision than they would if they had made the same decision at work [30]. Contextual factors that play into an individual's values may be factors specific to a country (average wealth, income inequality or equality, employment rates) [38].

2.2 Memory for conversational agents

A significant amount of research has focused on integrating memory into conversational agents and social robots [5]. Many of these designs are inspired by how human memory works. Human memory is generally represented by distinguishing between long-term, short-term and working memory on the one hand [1, 12, 17] and episodic and semantic memory on the other [15, 35, 39]. Most work within the human-agent and human-robot community has focused on modelling episodic memory. The episodic memory captures previous experiences embedding them within the context of time, location, and possibly the emotions experienced. A specific instance of an episodic memory is an autobiographical memory. This information has often been represented by at least *4 W*: *What, Where, Who* and *When* [10, 18, 20, 28].

Especially within the last two years, there has been a notable surge in efforts to equip large language models with memory capabilities. Most of these initiatives involve creating an agent's memory by supplying them with some form of structured prompt [22, 26, 34]. However, prompt-based methods have limitations as it becomes increasingly challenging to capture all information within a single prompt without introducing conflicting information, particularly as the history grows longer.

Behaviours and actions are interpreted at a surface level without modelling the underlying motivations and beliefs. While this method may yield impressive short-term results in terms of perceived intelligence, the failure to accumulate knowledge about interpreting user motivations and beliefs over time constrains the interpretive abilities of future actions.

Memory models so far have been used in an array of different use-case scenarios like education [16] and elderly care [7, 33].

3 METHOD

We designed a **value-aware memory** system to rank participants' values in a conversation. The contextual value memory attempts to rank what values are important for an individual user by mapping each value in Schwartz's value model (see Section 2.1) to a number. Each value in Schwartz's circle [29] is given a weight of 1 for each time the user chose that value when given the choice, and 0.25 for each time the user chose a value in the same quadrant of Schwartz's circle. Ties are broken by counting keywords in the users' verbal responses to the agent and mapping them to different values according to a translated variant of value-mapped dictionaries by Fischer et al. [14], Ponizovskiy et al. [27].

The value-aware memory was implemented in a *Nao* robot. We set up a between-subject experiment where the memory-equipped robot, supported by a screen showing images relating to the dialogue, talked to users about their values across three sessions. In the conversation, the participants selected options corresponding to specific values in response to prompting by the robot. The dialogue was built around **Socratic questioning** [6, 11, 25], in order to identify users' values in cooperation with them. The value model by Schwartz [29, 31] considers 10 categories of values. We instead limited the values explored in our dialogues to **five** of the ten to reduce the amount of dialogue that would have to be pre-written comparing pairs of values. Based on previous work by Suizzo [36], we decided to focus on **conformity**, **achievement**, **self-direction** and **benevolence**. **Hedonism** was also included to more evenly spread the values across the circle by Schwartz et al. [31], and to provide a value on the other side compared to *conformity* and *benevolence*. To also connect the experiment to previous work suggesting that values differ depending on the context where they are evaluated (see Section 2.1.1), we developed dialogues that evaluated participants' values in the three different contexts of *school*, *social* and *family*.

The three sessions were split across three days. In the first session, users were given two **scenarios** per context. Users were asked what they would do in the proposed situation, and they could respond with all five of the values we had chosen to include. In the second session, the same three contexts as in the first session were explored with two new scenarios each. For this session, the scenarios were evaluated by presenting the participant with two choices for each scenario – each choice represented a behaviour connected to a value in that scenario. In the third session, the robot pointed out aspects of the participants' value models where they had differing values depending on the context, and asked two sets of questions about what the participant thought of someone who would pick the opposite option of the one chosen by the participant in the second session, and what the participant would think about someone who valued a different value than one they chose during the first session.

There were two experimental conditions. In the **complete** condition, the system based the value-dependent parts of the second and third session on the user's previous responses. In the **partial** condition, the system made random choices whenever a scenario referred back to the users' previous value choices. 57 participants were recruited to participate. 28 (14 M, 14 F) were assigned to the *complete* condition, while 26 (14 M, 12 F) were assigned to the *partial* condition, with three participants excluded. The mean age was 38.3 years old, with a standard deviation of 18.8.

After the second and third session, structured interviews were performed to evaluate the relevance of the dialogue and the system. The questions from these interviews are presented in Appendices A.1 and A.2. After the third session, participants also filled in the **likeability** and **perceived intelligence** questionnaires from the Godspeed series [4], using the standardised Dutch translation [2]. Following this, two value models were presented visually to the participant in the form of three *spider charts*[1], one per context. Each chart contained graphs representing real and fake values side-by-side, with the fake values generated by shifting the real weights (created by the system through sessions 1 and 2) from their proper value into a different value. The participant then answered several questions about their impression of the two visualisations of the two value models, and their memories from the conversations with the agent throughout the three sessions.

4 RESULTS & DISCUSSION

Using a Mann-Whitney U test [24], **likeability** was found to not differ significantly ($U = 349$, $p \approx .79$) between the *complete* ($M = 4.3$, $SD = .5$) and *partial* ($M = 4.2$, $SD = .7$) conditions. **Perceived intelligence** was similarly found to not differ significantly ($U = 300.5$, $p \approx .27$) between the *complete* ($M = 4.0$, $SD = .7$) and *partial* ($M = 3.6$, $SD = 1.0$) conditions. Surprisingly, the absence of reliable value memory did not lead the participants to perceive the *partial* condition as less intelligent or less likeable than the *complete* condition. The between-subjects design meant that participants had to make up their own mind about the standard for how a *smart* agent behaves. One interpretation is that users thought of the robot's reasoning as somehow separate from the robot – as though the robot was static but that the job it was performing was different between the two conditions. It is possible that participants' responses to the Godspeed questionnaires related more to their impression of the robot's embodiment and speech style, which were thought of as separate from the dialogue design. Kasap and Magnenat-Thalmann [20, 21] found that an agent with episodic memory had a higher level of *social presence* than one without memory. We can conclude that *which* parts of memory that a human user perceives as crucial to the agent's social presence depends on the agent and dialogue setup. It is well-known that a robot's embodiment and human-likeness affects its perceived intelligence [3], so our usage of a clearly robotic NAO agent may have affected these results.

Participants' responses to three questions about the relevance of values discussed by the agent were found to correlate with a Cronbach's α [13] of 0.76. The answers to the three questions were thus averaged and compared between the two conditions. A Mann-Whitney U-test showed that the distribution in the *partial* condition ($M = 4.7$, $SD = .9$) was significantly different ($U = 176.5$, $p < 0.01$) from that in the *complete* condition ($M = 5.2$, $SD = .7$). A Mann-Whitney U test confirmed that the participants in the *complete* condition ($M = 5.4$, $SD = 1.0$) thought that the agent learned more ($U = 198$, $p < 0.005$) than the participants in the *partial* condition ($M = 4.5$, $SD = 1.0$). This implies that participants did perceive that the *partial* condition was not adapting to their values as much as the *complete* condition.

In the *partial* condition, 14 participants believed that the *fake* value model they were shown after the third session fit their values more than the *real* values generated from the system's value memory. 12 participants believed that the value memory was the better fit. In the *complete* condition, 27 out of 28 participants instead believed that the *real* values were a better fit. There was thus a strong significant relationship between the condition and which model participants chose ($\chi^2(1, N = 54) = 27.9$, $p < 0.001$). It is somewhat surprising that participants in the *partial* condition were not able to recognise their own values when shown the spider chart. The *real* values were based on user responses both in the *partial* and *complete* conditions. We presume that the difference comes from participants being affected by the system's random claims about their values during the third session.

[1] An example spider chart is shown in Appendix B, and the related questions are listed in Appendix A.2.

4.1 Differences between contexts

Repeated Friedman tests were performed per value, comparing the weight that the value had been assigned for each of the contexts explored in the dialogues (school, home, and social). The Friedman tests showed that the ranks were significantly differently distributed for conformity ($\chi^2(2, N = 54) = 24.3$, $p = 5.33 * 10^{-6}$), benevolence ($\chi^2(2, N = 54) = 9.89$, $p = 7.12 * 10^{-3}$) and self-direction ($\chi^2(2, N = 54) = 45.1$, $p = 1.63 * 10^{-10}$), while no significant differences were found for hedonism ($\chi^2(2, N = 54) = 7.51$, $p = 2.34 * 10^{-2}$) or achievement ($\chi^2(2, N = 54) = 8.93$, $p = 1.15 * 10^{-2}$).

To extract specific differences between the contexts, repeated Wilcoxon signed-rank post-hoc tests were run between all pairs of contexts for conformity, benevolence and self-direction. The tests found that **conformity** was valued more highly in the *home* context than in the *social* context ($U = 1064$, $p = 0.014$), more highly at *school* than at *home* ($U = 425$, $p = 0.00634$), and more highly at *school* than in the *social* context ($U = 110$, $p = 3.27 * 10^{-8}$). **Benevolence** was only found to differ such that it was more highly rated in the *home* than in a *social* context ($U = 1076.5$, $p = 0.0103$). **Self-direction** was valued more highly at *home* than at *school* ($U = 1134$, $p = 5.15 * 10^{-5}$), as well as more highly in a *social* context than at *school* ($U = 1477$, $p = 3.21 * 10^{-9}$).

5 CONCLUSIONS

Value-based memory models and user models can be useful for the future design of agents that assist users with specifically those hard questions where there is not one good, one-size-fits all answer. We can confirm that:

- **(i)** Building the robot's memory around participants' values was in fact noticeable to participants.
- **(ii)** Participants did perceive the value-aware condition as being aware of their contextual values, leading to more a value-aware conversation.
- **(iii)** Participants pripritised their values differently depending on the context.

ACKNOWLEDGMENTS

This research was funded by the Dutch-Swiss *ePartners4all* project (Grant no. TKI-LSH-T2019), and by the Dutch project *Technology Assisted Self-Management: Preventing Relapse and Crisis by the Severe Mentally Ill Themselves* (with project number KICH1.GZ03.21.002) of the research programme KIC - MISSIE 2021 which is (partly) financed by the Dutch Research Council (NWO). The authors would like to thank Deborah van Sinttruije for proofreading the final document.

REFERENCES

[1] Alan Baddeley. 2020. Short-term memory. In *Memory*. Routledge, London, UK, 41–69.
[2] Christoph Bartneck. 2023. Godspeed Questionnaire Series: Translations and Usage. In *International Handbook of Behavioral Health Assessment*, Christian U. Krägeloh, Mohsen Alyami, and Oleg N. Medvedev (Eds.). Springer International Publishing, Cham, 1–35. https://doi.org/10.1007/978-3-030-89738-3_24-1
[3] Christoph Bartneck, Takayuki Kanda, Omar Mubin, and Abdullah Al-Mahmud. 2009. Does the Design of a Robot Influence Its Animacy and Perceived Intelligence? *International Journal of Social Robotics* 1, 2 (01 Apr 2009), 195–204. https://doi.org/10.1007/s12369-009-0013-7
[4] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (01 Jan 2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3
[5] Paul Baxter and Tony Belpaeme. 2014. Pervasive memory: the future of long-term social HRI lies in the past. In *Third international symposium on new frontiers in human-robot interaction at AISB*. SSAISB, Bath, UK, 3 pages.
[6] Aaron T. Beck and David J.A. Dozois. 2011. Cognitive Therapy: Current Status and Future Directions. *Annual Review of Medicine* 62, 1 (2011), 397–409. https://doi.org/10.1146/annurev-med-052209-100032 arXiv:https://doi.org/10.1146/annurev-med-052209-100032 PMID: 20690827.
[7] Timothy W. Bickmore, Lisa Caruso, Kerri Clough-Gorr, and Tim Heeren. 2005. 'It's just like you talk to a friend' relational agents for older adults. *Interacting with Computers* 17, 6 (2005), 711–735. https://doi.org/10.1016/j.intcom.2005.09.002
[8] Rachel Burrows, Peter Johnson, and Hilary Johnson. 2014. Influencing Behaviour by Modelling User Values: Energy Consumption. In *2nd International Workshop on Behaviour Change Support Systems, PERSUASIVE'2014*. CEUR, Aachen, 85–93.
[9] Rachel Burrows, Peter Johnson, and Hilary Johnson. 2015. Value Sensitive Design Approach to Influence Energy-use Behaviour. In *INTERACT 2015 Adjunct Proceedings : 15th IFIP TC.13 International Conference on Human-Computer Interaction 14-18 September 2015, Bamberg, Germany*, Tom Gross and Christoph Beckmann (Eds.). 15th IFIP TC.13 International Conference on Human-Computer Interaction, 2015, University of Bamberg Press, Bamberg, 547–554. https://doi.org/10.20378/irb-58495 https://fis.uni-bamberg.de/handle/uniba/58495.
[10] Joana Campos and Ana Paiva. 2010. MAY: My Memories Are Yours. In *Intelligent Virtual Agents*, Jan Allbeck, Norman Badler, Timothy Bickmore, Catherine Pelachaud, and Alla Safonova (Eds.). Springer, Berlin, Heidelberg, 406–412.
[11] Gavin I Clark and Sarah J Egan. 2015. The Socratic method in cognitive behavioural therapy: a narrative review. *Cognitive Therapy and Research* 39, 6 (2015), 863–879.
[12] Nelson Cowan. 2008. Chapter 20 What are the differences between long-term, short-term, and working memory? In *Essence of Memory*, Wayne S. Sossin, Jean-Claude Lacaille, Vincent F. Castellucci, and Sylvie Belleville (Eds.). Progress in Brain Research, Vol. 169. Elsevier, Amsterdam, 323–338. https://doi.org/10.1016/S0079-6123(07)00020-9
[13] Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 3 (01 Sep 1951), 297–334. https://doi.org/10.1007/BF02310555
[14] Ronald Fischer, Johannes Karl, Velichko Fetvadjiev, Adam Grener, and Markus Luczak-Roesch. 2022. Opportunities and Challenges of Extracting Values in Autobiographical Narratives. *Frontiers in psychology* 13 (2022), 20 pages.
[15] Daniel L Greenberg and Mieke Verfaellie. 2010. Interdependence of episodic and semantic memory: Evidence from neuropsychology. *Journal of the International Neuropsychological Society* 16, 5 (2010), 748–753. https://doi.org/10.1017/S1355617710000676
[16] Helen Hastie, Mei Yii Lim, Srini Janarthanam, Amol Deshmukh, Ruth Aylett, Mary Ellen Foster, and Lynne Hall. 2016. I Remember You! Interaction with Memory for an Empathic Virtual Robotic Tutor. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (Singapore, Singapore) *(AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 931–939.
[17] G. J. Hitch and A. D. Baddeley. 1976. Verbal reasoning and working memory. *Quarterly Journal of Experimental Psychology* 28, 4 (1976), 603–621. https://doi.org/10.1080/14640747608400587
[18] Wan Ching Ho, João Dias, Rui Figueiredo, and Ana Paiva. 2007. Agents that remember can tell stories: integrating autobiographic memory into emotional agents. In *Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems* (Honolulu, Hawaii) *(AAMAS '07)*. Association for Computing Machinery, New York, NY, USA, Article 10, 3 pages. https://doi.org/10.1145/1329125.1329138
[19] Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. 1991. *Cultures and organizations: Software of the mind*. Vol. 1. McGraw-Hill, New York.
[20] Zerrin Kasap and Nadia Magnenat-Thalmann. 2010. Towards episodic memory-based long-term affective interaction with a human-like robot. In *19th International Symposium in Robot and Human Interactive Communication*. IEEE, New York, 452–457. https://doi.org/10.1109/ROMAN.2010.5598644
[21] Zerrin Kasap and Nadia Magnenat-Thalmann. 2012. Building long-term relationships with virtual and robotic characters: the role of remembering. *The Visual Computer* 28, 1 (01 Jan 2012), 87–97. https://doi.org/10.1007/s00371-011-0630-7
[22] Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. 2023. Prompted LLMs as Chatbot Modules for Long Open-domain Conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4536–4554. https://doi.org/10.18653/v1/2023.findings-acl.277
[23] Enrico Liscio, Michiel van der Meer, Luciano C. Siebert, Catholijn M. Jonker, Niek Mouter, and Pradeep K. Murukannaiah. 2021. Axies: Identifying and Evaluating Context-Specific Values. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (Virtual Event, United Kingdom) *(AAMAS '21)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 799–808.
[24] H. B. Mann and D. R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* 18, 1 (1947), 50 – 60. https://doi.org/10.1214/aoms/1177730491
[25] Christine A Padesky. 1993. Socratic questioning: Changing minds or guiding discovery. In *A keynote address delivered at the European Congress of Behavioural and Cognitive Therapies, London*, Vol. 24. EABCT, Tübingen, Germany, 6 pages.
[26] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (, San Francisco, CA, USA,) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. https://doi.org/10.1145/3586183.3606763
[27] Vladimir Ponizovskiy, Murat Ardag, Lusine Grigoryan, Ryan Boyd, Henrik Dobewall, and Peter Holtz. 2020. Development and validation of the personal values dictionary: A theory–driven tool for investigating references to basic human values in text. *European Journal of Personality* 34, 5 (2020), 885–902.
[28] Bart Schreuder Goedheijt. 2017. *Recalling shared memories in an embodied conversational agent: personalized robot support for children with diabetes in the PAL project*. Master's thesis. University of Twente.
[29] Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*. Vol. 25. Elsevier, Amsterdam, 1–65.
[30] Shalom H. Schwartz. 1997. *Values and culture*. Routledge, New York, NY, US, 69–84.
[31] Shalom H. Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, Ozlem Dirilen-Gumus, and Mark Konty. 2012. Refining the theory of basic individual values. *Journal of Personality and Social Psychology* 103, 4 (2012), 663–688. https://doi.org/10.1037/a0029393
[32] Ralf Schwarzer and Matthias Jerusalem. 1995. Generalized Self-Efficacy scale. In *Measures in health psychology: A user's portfolio. Causal and control beliefs*, J Weinman, S Wright, and M Johnston (Eds.). nferNelson, Windsor, UK, 35–37.
[33] Juhi Shah, Ali Ayub, Chrystopher L. Nehaniv, and Kerstin Dautenhahn. 2023. Where is My Phone? Towards Developing an Episodic Memory Model for Companion Robots to Track Users' Salient Objects. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 621–624. https://doi.org/10.1145/3568294.3580160
[34] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., Red Hook, New York, 8634–8652. https://proceedings.neurips.cc/paper_files/paper/2023/file/1b44b878bb782e6954cd888628510e90-Paper-Conference.pdf
[35] Larry R. Squire and Stuart M. Zola. 1998. Episodic memory, semantic memory, and amnesia. *Hippocampus* 8, 3 (1998), 205–211. https://doi.org/10.1002/(SICI)1098-1063(1998)8:3<205::AID-HIPO3>3.0.CO;2-I
[36] Marie-Anne Suizzo. 2007. Parents' Goals and Values for Children: Dimensions of Independence and Interdependence Across Four U.S. Ethnic Groups. *Journal of Cross-Cultural Psychology* 38, 4 (2007), 506–530. https://doi.org/10.1177/0022022107302365
[37] Bart Teeuw, Ralf Schwarzer, and Matthias Jerusalem. 1994. Dutch Adaptation of the General Self-Efficacy Scale.
[38] Raül Tormos, Christin-Melanie Vauclair, and Henrik Dobewall. 2017. Does Contextual Change Affect Basic Human Values? A Dynamic Comparative Multilevel Analysis Across 32 European Countries. *Journal of Cross-Cultural Psychology* 48, 4 (2017), 490–510. https://doi.org/10.1177/0022022117692675
[39] Endel Tulving. 1972. *Episodic and semantic memory*. Academic Press, Oxford, England, xiii, 423–xiii, 423.

A POST-SESSION QUESTIONNAIRES

The following tables list which questions were asked after which session, and which type of response was accepted for each question.

The questionnaires have been translated into English to present them here; they were presented to the participants in Dutch. Not all questions were analysed in this paper.

A.1 Questionnaire after session 2

Question	Answer type
Did you feel like the agent discussed relevant values during the conversation?	7-Point Likert Scale
Did you feel the agent learned something about you during the conversation?	7-Point Likert Scale
What do you not agree with during this conversation?	Open Question
What do you think the agent got wrong during the conversation?	Open Question
How relevant were the questions at the end of the conversation to your values?	7-Point Likert Scale

Table 1: The questions asked after the second session.

A.2 Questionnaires after session 3

The Godspeed questionnaires [4] on *likeability* and *perceived intelligence* were used in their generally accepted Dutch translation [2] and are not presented here. Participants also filled in the GSES scales [32] in their official Dutch translation [37], which are also not presented here. The results of the GSES questionnaires are also not analysed in this paper.

Question	Answer type
Do you agree with this model†?	7-Point Likert Scale
Did you feel like the agent discussed relevant values during the conversation?	7-Point Likert Scale
Did you feel the agent learned something about you during the conversation?	7-Point Likert Scale
What do you not agree with when seeing this model†?	Open Question
What did you learn from the conversation?	Open Question
What did you learn from seeing the memory models?	Open Question
What did this model† get wrong about you?	Open Question
How much do you remember from the conversation?	7-Point Likert Scale
What do you remember from the conversation?	Open Question
How much did the robot help you to self-reflect?	7-Point Likert Scale
Did you talk to a smart or dumb robot?	Smart/dumb

Table 2: The questions asked after the third session.
†: This question was asked with reference to the memory model representation that the participant had chosen out of the two shown to them at the start of the questionnaire.

B EXAMPLE SPIDER CHART

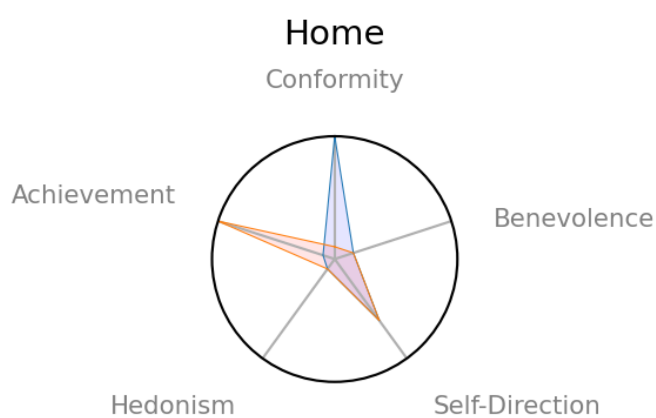

Figure 1: An example of a spider chart shown to the participant after the third session. Here, the blue graph is the participant's real values, with the red graph (partially overlapping) containing randomly shifted values.

Received 12 April 2024