Mind Modeling: A ToM-Based Framework for Personalization

Cristina Gena[1,*]

[1]Dept. of Computer Science, University of Turin, Italy

Abstract

User modeling has traditionally relied on inferring preferences, traits, or intents from observable behaviour. While effective in many adaptive systems, this paradigm treats behaviour as the primary object of modeling and leaves mental-state attribution implicit. This assumption becomes limiting in socially situated and longitudinal interaction, where behaviour must be interpreted in context and over time. We introduce *mind modeling*, a perspective in which user modeling is grounded in the explicit and revisable attribution of mental states, including beliefs, intentions, emotions, and knowledge. Drawing on Theory of Mind (ToM), this approach treats behaviour as evidence for hypotheses about internal states, supporting personalization that is more interpretable and coherent across interaction episodes. We present M^3, a conceptual framework that integrates perception, mentalisation, and action within a unified structure, enabling the continuous update of mental-state hypotheses in embodied interaction. We further illustrate this perspective through an embodied interaction trace, providing an initial operationalization of mind modeling in practice.

Keywords

User Modeling, Personalization, ToM, Mentalisation, Intentional Stance, Embodiment, Explainability

1. Introduction

User modeling has traditionally been framed as user profiling, i.e., the inference of user preferences, traits, or intents from observable behaviour and interaction histories [1, 2, 3]. This paradigm has enabled effective personalization in many adaptive systems, including recommender systems, intelligent tutoring systems, conversational agents, as well as adaptive mobile and in-car services [4, 5]. In its dominant form, profiling assumes that correlations between observed behaviour and inferred user characteristics are sufficient to support meaningful adaptation.

However, this assumption becomes increasingly fragile in embodied, socially situated, and longitudinal interaction, such as those involving conversational agents [6] or social robots [7], where behaviour is context-dependent, socially grounded, ambiguous, and requires interpretation over time [8]. In these settings, users interpret system behaviour in terms of beliefs, intentions, and emotions [9, 10, 11], and personalization is evaluated not only by predictive accuracy, but also by coherence, interpretability, and social appropriateness over time. The same observable behaviour may correspond to different underlying mental states (e.g., hesitation as uncertainty, anxiety, or avoidance), making purely correlational models difficult to interpret and justify.

To address this limitation, we introduce *mind modeling*, a perspective in which user modeling is grounded in the explicit and revisable attribution of mental states. Mind modeling treats behaviour as evidence for hypotheses about beliefs, intentions, emotions, and knowledge, which are continuously updated across interactions. This perspective is particularly suited to embodied settings where perception and action are tightly coupled and does not replace profiling-based approaches, but complements them: behavioural regularities and preferences remain, yet they are interpreted within a broader inferential process aimed at explaining why a user behaves in a certain way.

Joint Proceedings of the ACM UMAP Workshops 2026, UMAP 2026, June 8–11, 2026, Gothenburg, Sweden

*Corresponding author.

cristina.gena@unito.it (C. Gena)

https://www.di.unito.it/~cgena/ (C. Gena)

0000-0003-0049-6213 (C. Gena)

While the use of Theory of Mind (ToM) in artificial agents has been explored in AI and HCI [12, 13, 14], existing approaches are typically task-specific and do not frame ToM as a persistent, longitudinal user model for personalization. Our contribution is to position ToM as a foundational principle for user modeling itself, where mental-state attribution supports both adaptive behaviour and its explanation.

This paper makes three contributions: (i) it reconceptualizes user modeling as mental-state attribution grounded in interaction; (ii) it introduces M^3, a ToM-based conceptual framework integrating perception, mentalisation, and action; and (iii) it provides an initial operationalization through an embodied interaction trace and an implementation perspective, demonstrating how mental-state hypotheses can drive personalization and explanation over time.

2. Related Work

User modeling [1, 2], has long supported personalization in adaptive systems. Early approaches relied on symbolic and rule-based representations, while later work introduced probabilistic and data-driven methods, including Bayesian models [15], hidden Markov models [16], collaborative filtering [17], and deep learning techniques [18, 19]. These approaches have significantly improved scalability and predictive performance.

Despite their methodological diversity, most approaches share a correlational logic: they infer preferences, traits, or goals from observed behaviour without explicitly modeling the underlying mental states that generate such behaviour. As a result, adaptation is often driven by patterns in the data rather than by an explicit account of beliefs, intentions, or emotions. This limitation becomes more evident in socially situated and longitudinal interaction, where behaviour is context-dependent, ambiguous, and requires interpretation over time.

These challenges are particularly visible in embodied AI and social robotics, where systems operate in shared physical and social environments and interact through multiple modalities. In this context, user modeling must account not only for preferences, but also for situated action, affect, and interaction dynamics. Existing frameworks, such as the three-layer model by Rossi et al. [20], distinguish physical, cognitive, and social dimensions of user profiling, supporting tasks such as intention recognition, personality inference, and emotion detection. However, these approaches remain largely reactive and task-specific, and do not provide a persistent, revisable representation of the user as an intentional agent.

Theory of Mind (ToM), defined as the capacity to attribute mental states to oneself and others [21], provides a principled foundation for addressing this gap. ToM has been extensively studied across disciplines [22, 23], and includes both cognitive and affective components, supporting reasoning about beliefs, intentions, and emotions. There is no consensus on how humans implement Theory of Mind. Theory-Theory posits that individuals rely on an implicit folk-psychological theory to infer others' mental states [24, 21]. Simulation Theory instead suggests that people understand others by internally simulating their possible mental processes [25, 26]. In contrast, direct-perception and interaction-based approaches argue that some mental states can be directly perceived in embodied interaction, without full inferential reconstruction [27, 28]. Rather than committing to a single account, the framework proposed here is intentionally compatible with these perspectives. Theory-like components motivate explicit mental-state representation; simulation motivates the role of self-related representational resources in modeling others; and direct-perception highlights the importance of embodied and interactional cues, cautioning against treating all mental-state attribution as purely inferential. In this sense, M^3 can be interpreted as a unifying framework in which mental-state attribution is grounded in perception, supported by internal representational resources, and continuously refined through interaction.

Recent work in AI and HCI has explored ToM-inspired mechanisms in artificial agents [12, 13, 14, 29]. However, these approaches are typically task-specific or focused on agent interaction, and do not frame ToM as the basis of a persistent user model for personalization. In contrast, this paper proposes to treat user modeling itself as a process of explicit and revisable mental-state attribution, where behaviour is interpreted as evidence and personalization is grounded in the evolving representation of the user's

mind.

3. From User Modeling to Mind Modeling

In this paper, we define *mind modeling* as a form of ToM-based user modeling that maintains explicit, revisable hypotheses about user mental states, including beliefs, goals, intentions, emotions, and knowledge. These hypotheses are continuously updated on the basis of interaction evidence and are used to guide both personalization decisions and their explanation. Rather than treating behaviour as the primary object of modeling, mind modeling treats behaviour as the observable manifestation of underlying mental states that evolve across interaction episodes. This shift entails three consequences.
Representational consequence. Mind modeling represents users through structured models of mental states that are dynamic, uncertain, and context-dependent. User preferences and behavioural regularities are not discarded, but repositioned as evidence or secondary descriptors rather than the core ontology of the user model.
Inferential consequence. Personalization moves from direct mapping over behavioural traces to interpretative reasoning, where the same behaviour may support multiple competing hypotheses. Instead of assuming that one pattern implies one stable user characteristic, the system maintains alternative mental-state explanations and revises them as evidence accumulates.
Action consequence. Personalization becomes not only reactive but anticipatory and coherent over time. The system can justify not only what it did, but why it did it, by referencing the hypotheses that grounded the decision and their associated uncertainty.

This is precisely where mind modeling differs from both standard user modeling and conventional belief tracking. Probabilistic user models often include latent variables, but these are usually task-dependent abstractions introduced to improve behavioural prediction. In contrast, mind modeling treats beliefs, goals, and emotions as ontologically explicit mental-state hypotheses introduced to explain behaviour, not merely predict it. Moreover, belief tracking is often episodic and local, whereas mind modeling is intrinsically longitudinal, supporting revision and continuity across interaction episodes.

3.1. M^3: A ToM-based Framework

M^3 is a theory-based framework that specifies the representational and inferential commitments required for mind modeling. M^3 can be naturally interpreted within an embodied perception-action loop, where mental state attribution is grounded in interaction. In this perspective, mind modeling is not a static inference process but a continuous cycle in which the system perceives the user, updates hypotheses about the user's mental states, and produces situated actions that, in turn, shape future interaction. M^3 thus operates as the cognitive layer of an embodied agent, coupling perception, mentalisation, and action over time.

The framework introduces two coupled representational spaces: (i) an internal space encoding the system's own interaction-relevant states, and (ii) a user-directed space representing hypothesized user mental states with associated uncertainty. This distinction reflects the metacognitive nature of ToM, where attributing mental states to others presupposes internal representational resources.

In embodied settings, interaction evidence is grounded in multimodal perception, see for details [30, 31]. Let o_t denote raw observations at time t, including language, visual cues (e.g., gaze, facial expression), and behavioral signals (e.g., hesitation, actions). These observations are mapped to structured evidence:

$$E_t = \text{Percept}(o_t)$$

which serves as input to the mind modeling process.

The user mind model is then updated through a mentalisation operator:

$$m_{t+1} = U(m_t, E_t, s_t)$$

where m_t represents current hypotheses about user mental states, E_t is perceptual evidence, and s_t denotes the system's internal state.

Personalization is realized through situated action:

$$a_t = \pi(m_t, s_t, c_t)$$

where a_t corresponds to embodied behaviors such as speech or gesture and c_t encodes contextual constraints.

This process defines a closed loop: system actions affect the environment and the user, generating new observations o_{t+1}, which in turn produce new evidence E_{t+1}. As a result, mental state hypotheses are continuously revised in response to interaction dynamics.

M^3 can be described through three core elements.

Mental state space. Let M be a structured space of mental state variables (beliefs, goals, intentions, emotions, knowledge). The user model at time t is a state $m_t \in M$ with associated uncertainty.

Perception and evidence. Raw observations o_t are mapped to interaction evidence E_t, capturing signals relevant to mental state attribution.

Update and action. The system revises its hypotheses through an update operator U, and selects actions through a policy π, grounding decisions in explicit mental state representations.

This formulation highlights that, in M^3, behavior is not predicted directly from data but interpreted through hypothesized mental states that evolve over time. Embodiment ensures that these hypotheses remain anchored to perceptual signals and interaction context, enabling adaptive and explainable behavior in situated settings.

To illustrate how M^3 operates in practice, consider an assistive robot interacting with an elderly user during a daily medication routine.

t_0 - **Initial state.** The system maintains uncertain hypotheses about the user's beliefs regarding medication intake (e.g., whether the user believes the medication has already been taken).

Perception. The robot observes that the user does not mention the medication during a routine check-in and exhibits slight hesitation:

$$o_t \rightarrow E_t = \{\text{omission}, \text{hesitation}\}$$

Mentalisation update. Rather than inferring a generic state (e.g., low adherence), the system updates competing hypotheses:

- h_1: the user believes the medication was already taken
- h_2: the user is uncertain or has forgotten
- h_3: the user is avoiding the topic

Each hypothesis is associated with a confidence value and revised as evidence accumulates.

Action. The system selects a context-sensitive action:

$$a_t = \pi(m_t, s_t, c_t)$$

For instance, it may ask: *"I might be mistaken, but I was wondering if you already took your medication today?"*

Loop closure. The user's response generates new observations o_{t+1}, leading to further updates of the mental state model. This trace highlights the key difference from profiling-based approaches: behavior is not treated as direct evidence of a fixed trait, but as ambiguous data requiring interpretation. Embodied interaction provides the signals that ground this interpretation, while M^3 structures how these signals are transformed into revisable mental state hypotheses and actionable decisions. This abstraction provides the basis for concrete architectural instantiations, as outlined next.

3.2. Implementation Perspective

Building on these theoretical premises, we are currently implementing the approach within a hybrid cognitive architecture inspired by the Common Model of Cognition [32], adopting memory and control

mechanisms consistent with architectures such as SOAR, and extending the architecture with ToM as a new pillar for socially credible and personalized interaction. The proposed cognitive architecture integrates symbolic reasoning, subsymbolic perception, and embodied simulation, supported by SOAR-inspired memory subsystems. Dual knowledge sources, namely large language models with Retrieval-Augmented Generation (RAG) and declarative ontologies, jointly ensure scalability, contextualisation, and transparency of reasoning. At its core, a User Modeling Reasoner operationalises the Attribution of Mental States, constructing probabilistic user models from multimodal signals. This architecture instantiates the two representational spaces introduced in M3: an internal state space supporting reasoning and interaction management, and a user-directed space maintaining probabilistic hypotheses about the user's mental states.
Memory subsystems.

- **Working Memory** maintains the current interaction state and live ToM predicates (e.g., *user-believes(p)*, *user-feels(e, intensity)*).
- **Procedural Memory** stores policies governing sensorimotor routines and higher-level interaction strategies for prosocial behaviour.
- **Semantic (Declarative) Memory** encodes ontologies of mental states and domain knowledge, ensuring explicit and transparent representations [33].
- **Episodic Memory (Bayesian)** encodes interaction experiences and supports belief revision through Bayesian updating [34].

Dual knowledge sources.

1. **LLMs with RAG**, supporting dialogue and explanation through contextualised knowledge retrieval.
2. **Declarative ontologies**, enabling traceable and interpretable reasoning.

Cognitive layers.

- **Symbolic Reasoning Layer**: performs ToM reasoning, belief revision, and perspective-taking.
- **Subsymbolic Perception Layer**: extracts features from multimodal input using deep learning techniques.
- **Embodied Simulation Layer**: supports synchrony, mirroring, and affective resonance.

Bayesian episodic loop. A central mechanism integrates perception, reasoning, and memory through continuous updating of probabilistic user models, enabling both short-term adaptation and long-term personalization.
User modeling module. Within this architecture, a central component is the User Modeling module, which operationalises mind modeling by maintaining and updating mental-state hypotheses. It is inspired by the AMS framework [35], representing epistemic, emotive, intentional, imaginative, and perceptual dimensions with confidence scores in $[0, 1]$, structured as follows:

- **Input**: multimodal cues (e.g., facial expressions, gaze, prosody, posture, language).
- **Representation**: mental-state dimensions modeled as latent variables.
- **Confidence**: probabilistic estimates reflecting uncertainty.
- **Update**: belief revision through episodic memory.
- **Use**: adaptive behaviour and explanation generation.

In this setting, user preferences and behavioural regularities are not represented as primary model variables, but are treated as evidence derived from interaction history, contributing to the revision of mental-state hypotheses. This evidence is processed by the User Modeling Reasoner module, which maintains and updates a mental-state representation s with associated uncertainty. This representation is then used to guide behaviour selection and language generation, ensuring alignment between inferred user states and system outputs.

4. Conclusion and Future Work

Mind modeling reframes explainability as a structural property of personalization rather than an add-on. Because system behaviour is grounded in explicit mental-state hypotheses and their associated uncertainty, decisions can be traced and justified over time. This also supports longitudinal coherence, allowing the system to explain not only individual actions but also changes in behaviour across interaction episodes.

This perspective has implications for evaluation. While profiling-based systems are typically assessed through predictive accuracy, mind modeling requires additional dimensions, including: (i) coherence across episodes, (ii) plausibility of mental-state revision, and (iii) explanation helpfulness [36]. These dimensions can be operationalized through longitudinal analysis of state updates, agreement with human judgments or domain experts, and established HCI measures of clarity, trust, and perceived justification.

Mind modeling enables more adaptive and transparent interaction, but also introduces risks, including over-interpretation, privacy intrusion, and unwarranted confidence in inferred mental states. Systems may attribute beliefs or emotions that users have not explicitly expressed, or act on uncertain hypotheses in ways that influence user behaviour. These risks require explicit uncertainty representation, conservative update strategies, and transparency about what is inferred and why, together with mechanisms for users to understand and contest system assumptions.

This work represents an initial step toward operationalizing mind modeling and has several limitations. First, M^3 does not prescribe a unique computational implementation, and different realizations of U and π are possible. Second, ToM constructs are complex and their operationalization varies across domains and measures [21, 37]. Finally, while an initial implementation is underway, this paper does not provide empirical validation, but aims to establish a conceptual and architectural foundation for future work.

Ongoing work focuses on implementing and evaluating this approach in real interaction scenarios, assessing its impact on adaptation, explainability, and user trust in socially situated settings such as assistive robotics.

Declaration on Generative AI

The author employed ChatGPT exclusively for orthographic and grammatical proofreading of the final manuscript draft. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] A. Kobsa, Generic user modeling systems, User Modeling and User-Adapted Interaction 11 (2001) 49–63. doi:10.1023/A:1011140625715.

[2] P. Brusilovsky, E. Millán, User models for adaptive hypermedia and adaptive educational systems, in: P. Brusilovsky, A. Kobsa, W. Nejdl (Eds.), The Adaptive Web: Methods and Strategies of Web Personalization, volume 4321 of *Lecture Notes in Computer Science*, Springer, 2007, pp. 3–53. doi:10.1007/978-3-540-72079-9_1.

[3] E. Purificato, L. Boratto, E. W. De Luca, User modeling and user profiling: A comprehensive survey, arXiv preprint arXiv:2402.09660, 2024. doi:10.48550/arXiv.2402.09660.

[4] I. Amendola, F. Cena, L. Console, A. Crevola, C. Gena, A. Goy, S. Modeo, M. Perrero, I. Torre, A. Toso, Ubiquito: A multi-device adaptive guide, in: S. A. Brewster, M. D. Dunlop (Eds.), Mobile Human-Computer Interaction - Mobile HCI 2004, 6th International Symposium, Glasgow, UK, September 13-16, 2004, Proceedings, Lecture Notes in Computer Science, Springer, 2004, pp. 409–414. URL: https://doi.org/10.1007/978-3-540-28637-0_47. doi:10.1007/978-3-540-28637-0_47.

[5] C. Gena, I. Torre, The importance of adaptivity to provide onboard services: A preliminary

evaluation of an adaptive tourist information service onboard vehicles, Appl. Artif. Intell. 18 (2004) 549–580. URL: https://doi.org/10.1080/08839510490463442. doi:10.1080/08839510490463442.
[6] M. McTear, Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots, Springer Nature, 2022.
[7] C. Breazeal, K. Dautenhahn, T. Kanda, Social robotics, in: B. Siciliano, O. Khatib (Eds.), Springer Handbook of Robotics, Springer, 2016, pp. 1935–1972.
[8] A. Rapp, F. Cena, C. Gena, A. Marcengo, L. Console, Using game mechanics for field evaluation of prototype social applications: a novel methodology, Behaviour & Information Technology 35 (2016) 184–195. doi:10.1080/0144929X.2015.1046931.
[9] A. Marchetti, F. Manzi, S. Itakura, D. Massaro, Theory of mind and humanoid robots from a lifespan perspective, Zeitschrift für Psychologie 226 (2018) 98–109. doi:10.1027/2151-2604/a000326.
[10] D. C. Dennett, The Intentional Stance, MIT Press, 1987.
[11] A. Wykowska (Ed.), Intentional Stance toward Humanoid Robots: Lessons Learned from Studies in Human–Robot Interaction, Springer, 2024. doi:10.1007/978-3-031-55836-9.
[12] F. Cuzzolin, A. Morelli, B. Cirstea, B. J. Sahakian, Knowing me, knowing you: Theory of mind in AI, Psychological Medicine 50 (2020) 1057–1061. doi:10.1017/S0033291720000835.
[13] Q. Wang, A. K. Goel, Mutual theory of mind for human-AI communication, arXiv preprint arXiv:2210.03842 (2022). URL: https://arxiv.org/abs/2210.03842. arXiv:2210.03842.
[14] Q. Wang, S. Walsh, M. Si, J. Kephart, J. D. Weisz, A. K. Goel, Theory of mind in human-AI interaction, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 1–6. doi:10.1145/3613905.3636308.
[15] I. Zukerman, D. W. Albrecht, Predictive statistical models for user modeling, User Modeling and User-Adapted Interaction 11 (2001) 5–18. doi:10.1023/A:1011149806361.
[16] E. Horvitz, J. Breese, D. Heckerman, D. Hovel, K. Rommelse, The lumière project: Bayesian user modeling for inferring the goals and needs of software users, in: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI), 1998, pp. 256–265.
[17] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer 42 (2009) 30–37. doi:10.1109/MC.2009.263.
[18] F. Sun, Y. Guo, W. Zhang, D. Liu, Deep learning for intent recognition with multi-intent user utterances, IEEE Transactions on Knowledge and Data Engineering 33 (2019) 2236–2250. doi:10.1109/TKDE.2019.2940295.
[19] S. Zhang, L. Yao, A. Sun, Y. Tay, Deep learning based recommender system: A survey and new perspectives, ACM Computing Surveys 52 (2019) 1–38. doi:10.1145/3285029.
[20] S. Rossi, F. Ferland, A. Tapus, User profiling and behavioural adaptation for hri: A survey, Pattern Recognition Letters 99 (2017) 3–12. doi:10.1016/j.patrec.2017.03.002.
[21] H. M. Wellman, Making Minds: How Theory of Mind Develops, Oxford University Press, 2014.
[22] R. Saxe, N. Kanwisher, People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind", NeuroImage 19 (2003) 1835–1842. doi:10.1016/S1053-8119(03)00230-1.
[23] H. M. Wellman, Theory of mind: The state of the art, European Journal of Developmental Psychology 15 (2018) 728–755. doi:10.1080/17405629.2018.1435413.
[24] A. Gopnik, H. M. Wellman, Why the child's theory of mind really is a theory, 1992.
[25] A. I. Goldman, Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading, Oxford University Press, 2006.
[26] R. M. Gordon, Simulation without introspection or inference from me to you, Mental Simulation (1995).
[27] S. Gallagher, Direct perception in the intersubjective context, Consciousness and Cognition 17 (2008) 535–543.
[28] H. De Jaegher, E. Di Paolo, S. Gallagher, Can social interaction constitute social cognition?, Trends in Cognitive Sciences 14 (2010) 441–447.
[29] C. Gena, F. Manini, A. Lieto, A. Lillo, F. Vernero, Can empathy affect the attribution of mental

states to robots?, in: E. André, M. Chetouani, D. Vaufreydaz, G. M. Lucas, T. Schultz, L. Morency, A. Vinciarelli (Eds.), Proceedings of the 25th International Conference on Multimodal Interaction, ICMI 2023, Paris, France, October 9-13, 2023, ACM, 2023, pp. 94–103. URL: https://doi.org/10.1145/3577190.3614167. doi:10.1145/3577190.3614167.

[30] M. Mezzini, C. Limongelli, G. Sansonetti, C. De Medio, Tracking museum visitors through convolutional object detectors, in: Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '20 Adjunct, ACM, New York, NY, USA, 2020, pp. 352–355. doi:10.1145/3386392.3399282.

[31] A. Ferrato, C. Limongelli, M. Mezzini, G. Sansonetti, The meta4rs proposal: Museum emotion and tracking analysis for recommender systems, in: Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '22 Adjunct, ACM, New York, NY, USA, 2022, pp. 406–409. doi:10.1145/3511047.3537664.

[32] J. E. Laird, C. Lebiere, P. S. Rosenbloom, A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics, AI Magazine 38 (2017) 13–26. doi:10.1609/aimag.v38i4.2744.

[33] A. Lieto, Cognitive Design for Artificial Minds, Routledge, 2021.

[34] S. Vinanzi, M. Patacchiola, A. Cangelosi, Y. Demiris, Would a robot trust you? developmental robotics model of trust and theory of mind, Philosophical Transactions of the Royal Society B 374 (2019) 20180032. doi:10.1098/rstb.2018.0032.

[35] F. Manzi, G. Peretti, C. Di Dio, A. Cangelosi, S. Itakura, T. Kanda, H. Ishiguro, D. Massaro, A. Marchetti, A robot is not worth another: Exploring children's mental state attribution to different humanoid robots, Frontiers in Psychology 11 (2020) 2011. doi:10.3389/fpsyg.2020.02011.

[36] N. Tintarev, J. Masthoff, Designing and evaluating explanations for recommender systems, in: F. Ricci, L. Rokach, B. Shapira, P. B. Kantor (Eds.), Recommender Systems Handbook, Springer, 2011, pp. 479–510. doi:10.1007/978-0-387-85820-3_15.

[37] C. Osterhaus, S. L. Bosacki, Looking for the lighthouse: A systematic review of advanced theory-of-mind tests beyond preschool, Developmental Review 64 (2022) 101021. doi:10.1016/j.dr.2022.101021.