# Using Autoepistemic Logics for Understandable and Flexible User-Models

Johanna Wolff[1,*], Victor de Boer[2], Dirk Heylen[1] and M. Birna van Riemsdijk[1]

[1]*University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands*
[2]*Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands*

Abstract
Behavior change support agents are most effective when they are personalized to the user's goals and motivations. To achieve this the agent should be able to create a user model based on limited initial inputs from the user. We demonstrate how autoepistemic logic can be used to build a model which combines direct input from the user with assumptions about the user's reasoning. These beliefs can be used when reasoning about the user's motivations, but they may also be rejected when presented with conflicting information. This results in a user model in which both knowledge and beliefs about the user are included but still clearly separated. We illustrate our ideas using an example of a behavior support agent which assists the user in exercising more.

**Keywords**
Behavior support agent, User-Model, Shared mental models, Autoepistemic logic

## 1. Introduction

There is various technology that aims to support people who are in the process of changing their behavior or adopting new habits [1]. In order for these behavior support agents to effectively support the user, especially over a longer period of time, they need to be able to adapt to their user's goals, capabilities and preferences [2]. In this paper we use values to refer to the underlying reasons for choosing certain goals or actions [3]. This approach has been used in several systems [4], [5], especially because values are easily generalizable and tend to be relatively stable over time [6]. We take values to be the motivation for the goals that the user has set for themselves. Each action is connected to the goals it contributes towards or against and can either promote or demote a value. The values, goals and actions are each ordered by a priority relation which states how important they are to the user.

We see the agent and the user as a team and interpret the motivations of the user as a system which the agent and the user aim to optimize to achieve the goals of the user as much as possible. As described in [7], these teams can work together most effectively when they have a shared mental model of the system that is relevant to the task at hand. By representing the knowledge and the reasoning of the user model explicitly, we make it possible for the agent to explain to the user which information is being used, which initial theory that information is based on and which effects the information has on the agents output. This is also in line with a growing desire to ensure that artificial agents are designed responsibly and the user remains in control of how they use the technology [8]. An explainable agent can help the user understand and trust its suggestions [9], [10], [11].

If the user model is inaccurate, the user should be able to change the relevant information and adapt the agent to their needs. This may be the case because the reasoning of the agent was different than the users, information was missing or the user motivations change over time. Ideally, the agent is also able to recognize a conflict or gap in its knowledge base and ask the user for additional input to solve this. While the most accurate user model could theoretically be achieved by asking the user to input all details themselves, this would create a tedious user experience and deter people from engaging with the agent. Instead, the agent should be able to build a rich user model based off a few initial inputs by the user.

Human motivations can be incredibly complex since there are many different factors to consider when making a choice. The decision of whether someone wants to exercise can depend on the type of exercise, the time of day, the weather, and more. Additionally, there are many details which humans usually do not need to actively consider because they are not relevant. For example, it may be common to have few favorite sports but not to have a clear preference ranking of every sport. For these reasons, an agent's model of the user's motivations is unlikely to be perfectly accurate, especially considering the user's motivations often change over time. Therefore, instead

*21st International Workshop on Nonmonotonic Reasoning, September 2-4, 2023, Rhodes, Greece*

[*]Corresponding author.

j.d.wolff@utwente.nl (J. Wolff); v.de.boer@vu.nl (V. d. Boer); d.k.j.heylen@utwente.nl (D. Heylen); m.b.vanriemsdijk@utwente.nl (M. B. v. Riemsdijk)
0009-0005-0178-9570 (J. Wolff); 0000-0001-9079-039X (V. d. Boer); 0000-0003-4288-3334 (D. Heylen); 0000-0001-9089-5271 (M. B. v. Riemsdijk)

of focusing exclusively on the accuracy of the model, we emphasize the need for flexibility. Non-monotonic reasoning allows us to achieve this by making it easy to discard assumptions when new information contradicts them. We choose autoepistemic logic of knowledge and beliefs in particular because this allows us to treat the knowledge and the beliefs of the agent separately, which makes it easier to retrace where the information in the user model originates. This is especially useful when beliefs and knowledge contradict each other and we need to resolve the conflict. Additionally, by reasoning about the knowledge of the agent we can also express when something is not known and use this information to ask the user for additional input to build our model.

## 2. Autoepistemic Logic of Knowledge and Beliefs

We now sketch how autoepistemic logic of knowledge and beliefs can be used to build a flexible user model based on a few initial inputs from the user and assumptions by the agent. We separate these types of information by reasoning about both the agents knowledge and its beliefs and we base our framework on the autoepistemic logic of knowledge and belief developed in [12]. However, since we want to reason about different types of objects such as goals and values and the relations between them, we need to include first-order reasoning. We therefore use a first-order logic of knowledge an belief (*FOALKB*).

The language of *FOALKB* is a first-order modal language $\mathscr{L}_{K,B}$ with logical connectives $\vee, \wedge, \rightarrow, \neg, \bot$, quantifiers $\forall, \exists$, variables $x_i$, equality $=$, a set of predicate symbols $P_j$, a set of constants $c_l$ and modal operators $\mathscr{K}$ and $\mathscr{B}$ called knowledge and belief operators respectively. We allow arbitrary nestings of knowledge and belief operators, although they are not necessarily relevant to our application purposes. However, we do not allow the modal operators to be applied to formulas with open variables. Additionally, we restrict the quantifiers to formulas which do not contain any knowledge or belief operators. We are using constant domain semantics, which means we take our domain to be fixed in all expansions of our theory, so we can interpret a sentence of the form $\mathscr{B}\forall x P(x)$ to represent the set of sentences $\mathscr{B}p$ where $p$ is the proposition that expresses the truth value of $P(c)$ and $c$ ranges over all elements in our domain. Using this, we can translate all sentences from *FOALKB* into formulas of the propositional autoepsitemic logic of knowledge and beliefs introduced in [12]. For notation purposes we write our sentences in the language $\mathscr{L}_{K,B}$, but we use the propositional results obtained in [12].

We assume the following axioms and inference rules to describe the properties of knowledge atoms and belief atoms respectively.

($\mathrm{D_K}$), ($\mathrm{D_B}$) Consistency Axiom:

$$\neg\mathscr{K}\bot,\ \neg\mathscr{B}\bot$$

($\mathrm{K_K}$), ($\mathrm{K_B}$) Normality Axiom: for any sentences $F, G \in \mathscr{L}_{K,B}$

$$\mathscr{K}(F \rightarrow G) \rightarrow (\mathscr{K}F \rightarrow \mathscr{K}G),$$
$$\mathscr{B}(F \rightarrow G) \rightarrow (\mathscr{B}F \rightarrow \mathscr{B}G)$$

Knowledge and Belief Necessitation Inference Rule: for any sentence $F \in \mathscr{L}_{K,B}$

$$\frac{F}{\mathscr{K}F}\ ,\ \frac{F}{\mathscr{B}F}$$

The Consistency Axioms state that falsity is neither known nor believed. The Normality Axioms state that if $F$ implies $G$ is known (or believed) and $F$ is known (or believed) then $G$ must also be known (or believed). The Necessitation Rule expresses that everything that is provable in our logic is also known and believed.

In [12] the intended meaning of the belief operator is given by the condition that $F$ is believed in an expansion $T$ if $F$ is non-monotonically derivable from $T$:

$$T \vDash \mathscr{B}F \text{ if } T \vDash_{nm} F,$$

where $\vDash_{nm}$ denotes a specific non-monotonic inference relation. We will continue with the notion of minimal entailment which is also used in [12] which means that a sentence $F$ is believed to be true if it is true in all minimal models of the theory. A more in depth explanation can be found in [12].

For the knowledge operator $\mathscr{K}$ we use the interpretation

$$T \vDash \mathscr{K}F \text{ iff } T \vDash F,$$

which means that $F$ is known in an expansion $T$ if and only if $F$ is derivable from $T$.

When given an a *FOALKB* theory $T$, we are interested in the possible extensions. In our application, $T$ contains the initial inputs and the expansions of this theory will constitute our enriched user model. We want these expansions to be closed towards further reasoning which is referred to as a static autoepistemic expansion in [12].

We first define the set $Cn_*(T)$ as the closure of $T$, the smallest set which contains the theory $T$, all substitution instances of the axioms $\mathrm{D_K}$, $\mathrm{K_K}$, $\mathrm{D_B}$ and $\mathrm{K_B}$ and is closed under the necessitation rules and first-order logic. A static autoepistemic expansion is a theory $T^*$ which satisfies the following fixed-point equation:

$$T^* = Cn_*(T \cup \{\mathscr{K}F \mid T^* \vDash F\} \cup \{\neg\mathscr{K}F \mid T^* \nvDash F\}$$
$$\cup \{\mathscr{B}F \mid T^* \vDash_{\min} F\})$$

where $F$ ranges over all sentences in $\mathscr{L}_{K,B}$. In particular we are interested in the zero, one or several consistent static autoepistemic expansions of a theory.

The information in this model can be separated into three categories. The objective statements are are statements which are independent from the user, such as definitions of the objects and relations of the model. In our example we define the unary predicates Goal($x$), Value($x$) and Action($x$) to express that $x$ is a goal, a value or an action respectively, $\leq_G (x, y)$, $\leq_V (x, y)$ and $\leq_A (x, y)$ to denote priorities between goals, values and actions respectively, motiv($x, y$) to denote that a value $x$ motivates a goal $y$, adv($x, y$) to denote that an action $x$ contributes to achieving a goals $y$, prom($x, y$) to denote that an action $x$ positively relates to a value $y$, dem($x, y$) to denote that an action $x$ negatively relates to a value $y$ and their respective properties.

The knowledge of the agent comes from the direct inputs of the user. These sentences could take many forms but we give a few examples below.

$$\mathscr{K}\,(\text{Goal}(\text{GoForRun})) \tag{1}$$

This states that going for a run is a goal of the user.

$$\mathscr{K}\,(\text{Comfort} \leq_V \text{Social} \wedge \text{Social} \leq_V \text{Health}) \tag{2}$$

This expresses that the user prioritizes Health over Social Life and Social Life over Comfort.

$$\begin{aligned}\mathscr{K}\,(&\text{prom}(\text{GymFriend}, \text{Health})\\ &\wedge \text{prom}(\text{GymFriend}, \text{Social})\\ &\wedge \text{prom}(\text{Party}, \text{Social}))\end{aligned} \tag{3}$$

This expresses that going to the gym with a friend promotes the values Health and Social and going to a party promotes Social.

The beliefs of the agent are based on assumptions which the agent uses in its reasoning process. These assumptions may have been explicitly included during the design of the agent, based on previous data or psychological research, or they may be formulated during use of the agent based on current data regarding the user. We give some examples of beliefs which we may want to incorporate in our example agent.

$$\mathscr{B}\,(\forall x, y, z\,:\,(x \leq_V y \wedge y \leq_V z) \rightarrow x \leq_V z) \tag{4}$$

This expresses that the priorities the user has between values are transitive.

$$\begin{aligned}\mathscr{B}\,(\forall a, v\,:\,&\text{Action}(a) \wedge \text{Value}(v) \wedge \neg\mathscr{K}\text{prom}(a, v)\\ &\rightarrow \text{dem}(a, v))\end{aligned} \tag{5}$$

This expresses that if we do not know that an action promotes a value, then we assume that the action demotes the value instead. All these belief sentences express plausible assumptions in the context of our exercise support agent. While these are relatively simple examples, we can see that this framework provides us opportunities to infer additional assumptions which would normally not be included in the expansions of our theory.

If we take $T$ to be the set of all objective sentences and the knowledge sentence (2), we observe that Comfort $\leq_V$ Health $\notin Cn_*(T)$ since we have no information about the relation between Comfort and Health. In fact, the static expansion $T^*$ would even contain $\mathscr{B}\neg$(Comfort $\leq_V$ Health) if we use minimal entailment to interpret the belief operator. This would obviously not be useful for our user model. We could have avoided this situation by asking the user to provide a full ranking of the values, which would have been acceptable in this simplified scenario with only three different values. However, in more complex scenarios this is no longer feasible. No-one wants to provide an ordered list of their top 100 activities but they will probably be willing to provide their favorite or decide between two options. By including belief sentences such as (4) the expansion $T^*$ will now contain the belief sentence $\mathscr{B}$(Comfort $\leq_V$ Health), just as intended.

Next, we take $T$ to be the set of all objective sentences and knowledge sentences, but omit the belief sentences. In particular we are interested in how different actions relate to the values we have. Since we have no information about any actions demoting values, the static expansion $T^*$ would not only contain $\neg\mathscr{K}$dem(GymFriend, Comfort) but also $\mathscr{B}\neg$dem(GymFriend, Comfort). This may be warranted if we assume that the user would tell us if an action affects a value in any way and we accept that the value Comfort is not affected by going to the gym with a friend. However, if we include belief sentence (5), then we clearly see that $\mathscr{B}$dem(GymFriend, Comfort) $\in T^*$.

## 3. Conclusion

By using *FOALKB* we can build an enriched user model based on incomplete initial inputs from the user and assumptions which the agent has about the user model. In the created model we can easily distinguish between information which is based directly on knowledge about the user and information which the agent has infered based on other assumptions. In future work we will explore how this affects the understandability and trustworthiness of the agent. Additionally, we want the agent to allow for additional inputs from the user in case their motivations change or the model is inaccurate. We will explore how we can best incorporate knowledge and belief revision into our framework to make this possible. Lastly, we will explore which additional challenges arise when implementing the framework into a suitable logic programming language. This includes looking into the computational complexity and possibly placing additional restrictions on the logic.

## Acknowledgments

This research was funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, grant number 024.004.022.

## References

[1] H. Oinas-Kukkonen, Behavior change support systems: A research model and agenda, in: T. Ploug, P. Hasle, H. Oinas-Kukkonen (Eds.), Persuasive Technology, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 4–14. doi:10.1007/978-3-642-13226-1_3.

[2] M. B. van Riemsdijk, C. M. Jonker, V. Lesser, Creating socially adaptive electronic partners: Interaction, reasoning and ethical challenges, in: Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS '15, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2015, p. 1201–1206.

[3] B. Friedman, P. Kahn, A. Borning, P. Zhang, D. Galletta, Value Sensitive Design and Information Systems, 2006, pp. 55–95. doi:10.1007/978-94-007-7844-3_4.

[4] M. L. Tielman, C. M. Jonker, M. B. van Riemsdijk, What should i do? deriving norms from actions, values and context, in: MRC@IJCAI, 2018.

[5] T. L. van der Weide, F. Dignum, J. J. C. Meyer, H. Prakken, G. A. W. Vreeswijk, Practical reasoning using values, in: P. McBurney, I. Rahwan, S. Parsons, N. Maudet (Eds.), Argumentation in Multi-Agent Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 79–93.

[6] S. Schwartz, Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries, volume 25, 1992, pp. 1–65. doi:10.1016/S0065-2601(08)60281-6.

[7] C. Jonker, M. Riemsdijk, B. Vermeulen, Shared mental models - a conceptual analysis., in: Coordination, Organizations, Institutions, and Norms in Agent Systems VI - COIN 2010 International Workshops, 2010, pp. 132–151.

[8] S. S. Sundar, Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII), Journal of Computer-Mediated Communication 25 (2020) 74–88. doi:10.1093/jcmc/zmz026.

[9] S. Anjomshoae, A. Najjar, D. Calvaresi, K. Främling, Explainable agents and robots: Results from a systematic literature review, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2019, p. 1078–1088. doi:10.5555/3306127.3331806.

[10] A. B. Haque, A. N. Islam, P. Mikalef, Explainable artificial intelligence (xai) from a user perspective: A synthesis of prior literature and problematizing avenues for future research, Technological Forecasting and Social Change 186 (2023). doi:10.1016/j.techfore.2022.122120.

[11] S. Lockey, N. Gillespie, D. Holm, I. Asadi Someh, A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions, in: Proceedings of the 54th Hawaii International Conference on System Sciences, 2021, pp. 5463 –5472. doi:10.24251/HICSS.2021.664.

[12] T. C. Przymusinski, Autoepistemic logic of knowledge and beliefs, Artificial Intelligence 95 (1997) 115–154. URL: https://www.sciencedirect.com/science/article/pii/S0004370297000325. doi:https://doi.org/10.1016/S0004-3702(97)00032-5.