Prediction of Users Perceptional State for Human-Centric Decision Support Systems in Complex Domains through Implicit Cognitive State Modeling

Sergey V. Kovalchuk (sergey.kovalchuk@huawei.com)
Huawei
Saint Petersburg, Russia

Ashish Tara Shivakumar Ireddy (ireddy@itmo.ru)
ITMO University
Saint Petersburg, Russia

Abstract

This paper presents an approach to model the internal cognitive state of decision-makers when interacting with AI to understand exchanges between agents and improve future interactions. We focus on understanding how AI suggestions are perceived by a human agent using an approach based on the technology acceptance model. The variation in the user's state is investigated when perceiving the interaction with AI by considering it as a hidden (latent) state. Using human evaluation data collected from two cases of clinical decision-making and software development scenarios, we analyse and explore the user's perceptional state during interaction. The experiment conducted employs the Bayesian belief network to represent the human perceptional model and provide a prediction of the usefulness of AI model's suggestions in the considered case. Upon introduction of cognitive states in the model, we observed an increase in predictive performance by 76–77%. Our investigation can be concluded as an attempt to identify implicit static and dynamic cognitive characteristics of users to provide personalized assistance in human-AI interaction (HAI) and collaboration in complex domains of decision-making.

Keywords: human-AI interaction; human perceiving; cognitive states; decision support systems

Introduction

With recent advances, artificial intelligence (AI) has expanded its outreach into multiple areas including many critical sectors such as medicine, autonomous vehicles, defense technologies etc. The evolution of an AI model is based on its interaction with the environment and its respective users. With this in mind, human – AI interaction is a key issue within the community that aims to understand how an AI model can better serve the human user and further provide efficient and effective results. Yet, the development of human-AI interaction (HAI) is still an open issue. One of the ways to address this issue is with the introduction of guidelines and various regulatory directives (Amershi et al., 2019) that structure the way in which interactions are managed. Nevertheless, the problem has many aspects to it. One of the leading issues is related with attaining a proper understanding of human behaviour (e.g. decisions and choices) and its relevant reasoning. This issue is currently approached through techniques such as inverse reinforcement learning, theory of mind etc (Howes, Jokinen, & Oulasvirta, 2023; Jara-Ettinger, 2019). Parallelly, the other side of this problem is often overlooked. It deals with providing a better understanding of AI-generated results to humans via explainability, interactive analysis etc (Sreedharan, Kulkarni, Smith, & Kambhampati, 2021; Maclure, 2021), as a way to improve trust and further enhance the model itself. In order to develop a proper human – AI interaction standard there is a need for accurate alignment of information, evaluated by human agents and AI models alike. It is important in both Offline and Online modes of AI to obtain relevant metrics during the training procedure and acquire optimal feedback respectively. This is crucial in complex domains where significant expertise and background knowledge are required to evaluate results. In such domains, AI solutions (e.g. Applications in question-answer systems) that are relatively close to each other may significantly differ from one another concerning the problem being solved. Yet, distinct answers may be more accurate than the rest. A series of works focused on this sphere recently appeared being aimed at developing domain-specific quality metrics (see for example domain-specific metrics in medicine (Taha & Hanbury, 2015), programming (Evtikhiev, Bogomolov, Sokolov, & Bryksin, 2023) and other domains). However, the developed metrics are usually rather narrow and limited as they tend to be centred on aspects relevant to individual models or specific domain data. Thus, many recent studies rely on human feedback for training models (Nakano et al., 2021). As a result, the state of humans perceiving information from AI is considered a complicated problem to be addressed.

Within this study, we propose an approach towards modelling "human perception" during HAI. A particular interest of this study is in overseeing AI-based decision support recommendations in complex domains that require significant expertise to make decisions. By observing this area as a large multi-disciplinary field, we propose and evaluate a core conceptual approach for analysis and prediction of internal human reasoning while taking into account the multi-dimensional nature of a decision maker's perceptional and implicit cognitive states (i.e. the thought process during decision making relative to the environment and its interpretation). The paper reports early results in the investigation and analysis of the proposed approach. We believe the approach could be of interest to readers working in the area of HAI and human-centred evaluation of AI.

Background and related works

The architecture of HAI is based on the exchange of information between human agents and AI models. Each exchange has its respective significance and implications on the out-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

come. Present-day HAI models can be compared to legacy power systems or black box approaches. E.g., (Kazerooni, 1993) have used a hydraulic system to study and determine the ground rules for robotic system controls.

Human-to-human interaction can be an ideal scenario to define perceiving information between agents which can be done in many ways i.e. logically, mathematically, etc, with the end goal to extract a merit of significance that depicts a certain outcome to be produced by the model. The dual process theory (Evans & Stanovich, 2013) provides an ideal benchmark to view the emergence of a thought process based on prior knowledge and dynamic intuitive actions. With a majority of already existing architecture being a combination of either types, the viability of the exact thought process generated is left upon the decision-maker to recommend rather than the AI model itself. (Miller, 1956) highlights in his work that there is a limited chunk of information that can be retained as meaningful via an interaction depending on the timeline of the user's memory, which can be assumed to define the relative perceptional state and therefore the outcome as well. In these scenarios, the use of features (subjective, descriptive, psychological etc) can be crafted into a pipeline to generate the perceived thought process of the decision maker. A direct representation of such metrics would be defined using metrics such as trust, faith, competitiveness, background, training data etc, for the model to attain user validation to enhance the overall decision process. (Sundar, Kim, & Gambino, 2017) proposes that there exists a psychological mechanism for decrypting the information from users via the medium of communications where interactivity is considered a crucial pillar that can define the magnitude of impact and still be able to define boundaries of differences Theory of Interactive Media Effects (TIME). As information is exchanged in multiple modes of interactions the whole scenario put together provides certainty on the outcome of the interaction that can be modelled from either perspective, where both have a common goal and individual sub-goals relative to the environment.

HAI has many ecosystems per se. During our search, we found that dedicated systems such as chat-bots, virtual assistants, etc. have a fixed goal of answering limited questions. By combining concepts such as Theory of Mind (ToM) and Planned Behavior (TPB), Human In The Loop (HITL), etc., a complete process of interaction can be created such as the Technology Acceptance Model (TAM). Measuring perception or understandability of user input has to be paid attention to as it diverges from the use of general tools to the implementation of reinforcement learning, multi-agent reinforcement learning, behavioural cloning and other such approaches that aim to model the human input as a perspective answer rather than a preset based query.

Extended human-AI interaction

General approach to pipeline description

Here, we consider a comprehensive procedure of AI-assisted decision making (see Fig. 1). A basic decision making action performed by a human is based on the information related to a particular case and the environment ((1) in figure 1) processed by a decision maker. A common approach to the introduction of AI-assisted decision-making is usually focused on providing the (human user/decision maker) with case-specific reasoning and providing AI-generated claims (suggestions, plans, solutions, additional information). In most cases, the training of AI models is designed to "replicate" a Human decision-maker's behavior acting in similar historical cases. This is widely done using data-driven analysis of previously observed historical decisions ((2) in figure 1). Nevertheless, we see several ways to improve this training approach by extending the common pipeline in various aspects.

Yet, the important aspect of AI-assisted decision-making is the level of trust given by the user to AI claims. We see this procedure as an internal human evaluation procedure. This internal evaluation affects final decision-making sequence by taking into account provided AI claims. Many AI algorithms may provide specific assessments of the generated recommendation which may be measures of confidence, expected precision, internal uncertainty, etc. Additionally, this information can be further extended with explanations, semantic linking, relevant historical cases, etc., this expansion provides a more comprehensive view of the human agent. Despite this, the final evaluation of the result after decision making is usually performed by human users (decision-makers) and often involves many psychological aspects and professional backgrounds to assess whether the information provided by AI is correct and suitable. With this in mind, we suppose that the extended assessment procedure which takes into account internal human-centred evaluation could be a better source for AI training and internal assessment ((3) in figure 1). Moreover, it takes into account both the complexity of the problem domain and the personal characteristics of decision makers to make AI suggestions increasingly personalized. At the same time, internal cognitive context representation is a problem requiring further investigation and analysis.

Next, both human and AI agents have specific resources, limitations, and characteristics. Human decision-makers usually take into account domain-specific knowledge, professional experience, and cognitive skills, while AI models are dependent on their training environment and learning curve that impacts their behavior in recommending decisions. Furthermore, considering HAI, it is important to mention that professionals often vary in the way they see AI in general and AI's recommendations in particular scenarios. We can consider a wide range of factors influencing this procedure including subjective attitude towards AI, professional expertise, goals and values, etc. At the same time, AI agent is often limited in informational and computational resources and work to a known or unknown level of uncertainty. Additionally, recommendations, decisions, and evaluations generated by agents are not totally internal. The actions have effects on the environment and the particular case under consideration ((1) in figure 1) which should also be considered.

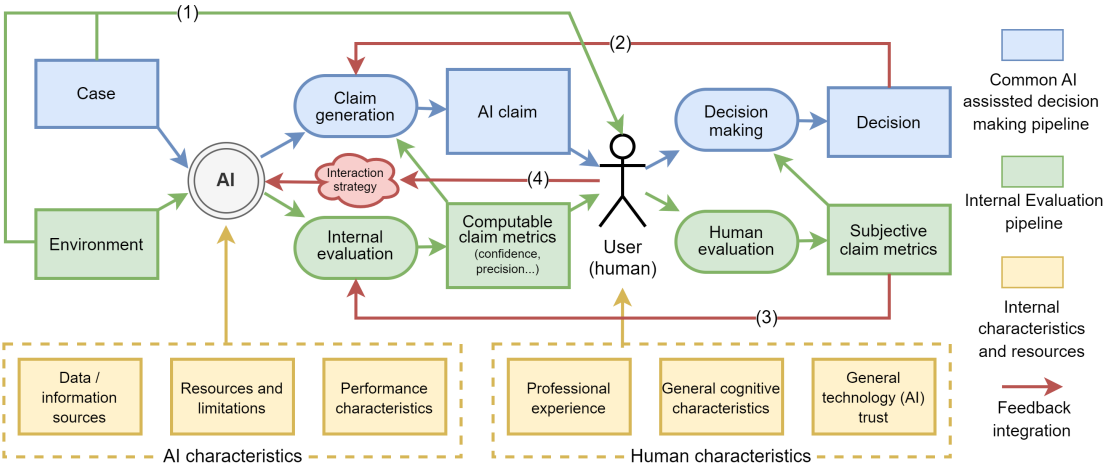

Figure 1: Overview pipeline of Human-AI interaction. Pathway (1) represents the Components involved in a basic decision, Pathway (2) represents the data driven analysis done using historical cases, Pathway (3) represents the extended assessment and Pathway (4) represents the human users perception to the AI model

Finally, an important aspect is explicit HAI strategies. We believe that modern AI should be beyond "replicating" human decision-makers. It should be considered as a collaborator with a specific role. Therefore the strategy for optimization should include subjective and objective values of collaborative (human+AI) decisions having conditions of both agents and communicating environment. Thus, an AI agent should have an understanding of how a "human partner" will perceive the recommendations (Howes et al., 2023) to optimize interaction strategy ((4) in figure 1).

Subjective assessment model in HAI

We can consider a decision process as a repetitive action A performed by a decision maker, given the case and environment state S (it can be considered as a Markov Decision Process, MDP). There exists a cognitive state C of the human user that affects internal cognitive metrics M_C derived from the current state S and AI claims I_{AI} (it also can be considered as part of the information system's state or observations, e.g. as $S^* = \langle S, I_{AI} \rangle$): $\mathcal{P}(M_C|C,S^*)$. Then, we consider metrics M_C affecting (implicitly or explicitly) expected decision-making reward $\mathcal{R}(s,s'|M_C)$ after changing the state from s to s' (which may be wrong but is used by a decision maker for selection strategy).

Given the fact that the cognitive state C of a human user can't be observed explicitly, there are two key problems in the implementation of the proposed scheme. First, structuring and identifying C such that it comprehensively describes the key characteristics of the human user. One of the ways to describe and identify crucial characteristics may be carried out via cognitive architectures (Kotseruba & Tsotsos, 2018). Nevertheless, the question of structural identification is still open. In particular, one can include a multitude of relevant characteristics that describe the human user such as:

- **Professional** characteristics (expertise, experience, etc.) including "hard" and "soft" skills;
- **Psychological** characteristics including general and AI-relatable characteristics (flexibility, attitude to new technologies, etc.);
- **Values and Goals** including directions in personal and professional life.

The second problem is related to building a proper interconnection between the human user's cognitive state C, the expected reward, and the relative actions being taken. This interconnection should take into account a multitude of objective characteristics that are specific to the respective domain and its environment. The list of such characteristics includes:

- External **environment** (physical, informational, etc.) and its characteristics affecting the decision process;
- **Case** level characteristics including the complexity of an individual case, its risks, uncertainty, etc.;
- **Role** of the decision maker themselves, i.e. professional position (occupation), level of responsibility, etc.

Nonetheless, a more sophisticated issue is the creation of a detailed structure between interconnections and rewards. One of the ways to consider this problem is inverse reinforcement learning approach to identify the reward function through observations that require a wider range of scenarios to be evaluated by human experts. In our case we consider extending the interconnection structuring through internal cognitive metrics M_C which can be evaluated separately through surveys, feedback analysis, etc., to thereby understand the procedure of implicitly perceiving the reasoning of experts during implementation.

Within our ongoing study, we consider a perception model constructed using three general scores (i.,e. applicable to multiple domains):

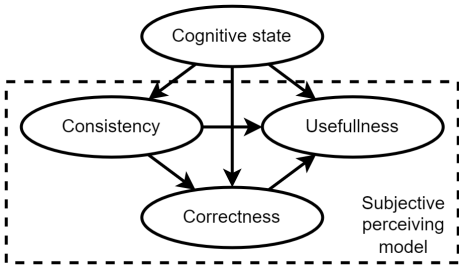

Figure 2: Experimental Bayesian belief network with cognitive state

- **Consistency.** Internal consistency of an AI claim including correct structure, absence of contradicting information, logical and lexical consistency, coherence, etc.

- **Correctness.** Case-specific correctness including relevance to the case or query, general domain-specific correctness, etc.

- **Usefulness.** Subjective assessment of a recommendation by the perceived intention to use the provided information in this particular case and environment.

The proposed list can be extended with more characteristics and evaluation metrics such as conciseness, completeness, justification, etc. (Breck et al., 2000). The scores and their interconnections can be evaluated through separate surveys with real or artificial data providing offline data for M_C evaluation. For example, we can use the Likert scale to quantify the metrics. Additionally, the structure and mutual influence of such metrics may be based on existing frameworks like TAM, and TBP (Dillon, 2006).

Within this study, we have focused on the listed characteristics for consistency and interpretability in various domains. Having an evaluation framework as such, we aim to consider the role of the internal cognitive state of the decision-maker during HAI. Fig. 2 shows a possible representation of a model in the form of a Bayesian belief network (BBN) with the explicit cognitive state of a user affecting perceiving scoring. Here we consider the connection of cognitive state C to the mentioned scores $\langle S_C, S_A, S_U \rangle$ where S_C denotes the evaluated internal consistency, S_A - agreement on the correctness of the solution, S_U - perceived usefulness.

The cognitive state variable C is a latent variable that can't be observed directly, however, it could be represented by a cognitive architecture, theory of mind, belief-desire-intention model or other approaches. Within our preliminary study, we implemented two basic approaches to represent the cognitive state to investigate the influence on prediction of M_C. The first approach is "one-hot" embedding for considering individual cognitive state constant to a user. Being rather limited in its interpretable representation of cognitive states, this approach enables the evaluation of personal cognitive characteristics which further could be mapped onto expertise, cognitive skills, and other subjective characteristics within the internal cognitive state model.

Table 1: Prediction of Usefulness score (S_U) with C classes of cognitive states

Dataset	*Samples*	*Users*	$MAE_0(S_U)$	C	$MAE_C(S_U)$
CDSS	570	19	0.4385	2	0.3132
				4	0.2440
				8	0.0965
CNLG	614	46	0.3208	2	0.1738
				4	0.1101
				8	0.0791

The second approach to the evaluation of cognitive state variables was to use BBN trained on one-hot embedding and use predicted probabilities in the latent space to obtain embeddings with the observed behaviour (i.e. given other variables). With this approach, we believe that we can observe variations in the behaviour of a user even within one decision maker and obtain more complicated patterns in latent variable space. For example, we can apply clustering to obtain a lower number of possible values of cognitive state variables and use it for further analysis, prediction and interpretation.

One of the possible ways is to use the information on cognitive state variables to predict personalized perceiving metrics and understand the user's state more precisely. The result would be a theoretical improvement in the human user's implicit acceptance of AI's claims therefore relatively better outcomes.

Experimental evaluation of cognitive state role

Case studies

In the course of this experimental study, we utilised two datasets collected during our previous experiments in the sphere of decision support and interaction analysis. Either case studies include a collection of feedback from domain specialists in corresponding areas with metrics considered in advance (consistency, correctness, usefulness).

Case 1. Clinical decision support system (CDSS) in type 2 diabetes mellitus (T2DM): In this study, an AI-based decision support system was implemented to provide risk prediction of patients who might be suffering from T2DM with various levels of explanation. The feedback was collected using a mock-up UI with synthetic patient data and risk prediction provided by a machine learning model measured via the risk assessment scale. Additionally, different levels of information were considered in the experimental study: a case when the AI model's prediction was provided exclusively, a case when the model's prediction was provided with the risk assessment by FINDRISC scale and a case when both model prediction and FINDRISC was provided along with additional explanation of the AI prediction. A significant varia-

tion factor here was the patient's condition (i.e. severity of the patient's status) and case complexity (presence of comorbidities, complications, etc.). During the experimental study, a total of 570 feedback results were collected from physicians of various specializations and experiences. The feedback includes scoring subjective understanding (consistency), agreement (correctness) and intention to use (usefulness) of the AI prediction. Additional information on the dataset collecting and processing can be found in (Kovalchuk, Kopanitsa, Derevitskii, Matveev, & Savitskaya, 2022).

Case 2. Code natural language generation (CNLG): In this study, a series of experiments were conducted using natural language generation via large language models (LLM). The task of this experiment was to answer programming-related questions with relevant source code. Several models (both originally trained by the publishers and versions fine-tuned by our team) were used to generate answers to questions that originated from StackOverflow[1]. The answers are generated as short code snippets in Python language and are compared with reference solutions (originally published at StackOverflow) using common natural language processing metrics by human assessors. During the experimental study, a set of 614 results generated by different LLMs were collected from 46 users working as software developers with different experiences (from students in software engineering to senior software developers). The feedback consists of the generated code that has undergone subjective evaluation for internal consistency (i.e to ensure that the answer provided is a piece of code without obvious errors), correctness (i.e. to ensure the result is answering the particular question), and usefulness (the prospect of usability of the results). Additional information on the dataset collecting and processing can be found in (Kovalchuk, Lomshakov, & Aliev, 2022).

The collected datasets were preprocessed to be represented in a single scale. In particular, a discrete 5-level Likert scale was used to represent each of the 3 answers of subjective evaluation metrics, where -2 relates to a "completely disagree" score and $+2$ relates to a "completely agree" score.

Implementation details

We have implemented the Bayesian belief network (BBN) as showcased in Fig. 2 using Pomegranate[2] library. For clustering and visual analysis of obtained embeddings, we utilized the Self-organized maps (SOM) approach as implemented in MiniSom[3] library. For identification of clusters and latent cognitive states, we have implemented BBN training using the leave-one-out cross-validation approach with the task of predicting the usefulness, S_U metric ("intention to use") as the latest scoring value. To evaluate the prediction, the mean average error ($MAE(S_U)$) calculated from the maximum probability score was selected as the target evaluation parameter. Using this target metric, we conducted a series of experiments within which several cluster variations were identified through the Self-organized map (SOM) approach within a range of $C \in [2,3,...,20]$ and an evaluation $MAE_C(S_U)$ was carried out for each dataset. To measure the influence of cognitive state introduction into the BNN, we have considered $MAE_C(S_U)$ in contrast to the prediction of S_U , without considering information on user identity and given only the other scores ($MAE_0(S_U)$).

[1] https://stackoverflow.com/
[2] https://github.com/jmschrei/pomegranate
[3] https://github.com/JustGlowing/minisom

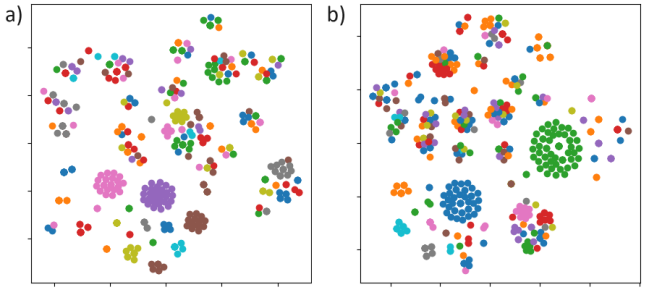

Figure 3: TSNE visualization of one-hot embeddings for various users in a) CDSS dataset; b) CNLG dataset

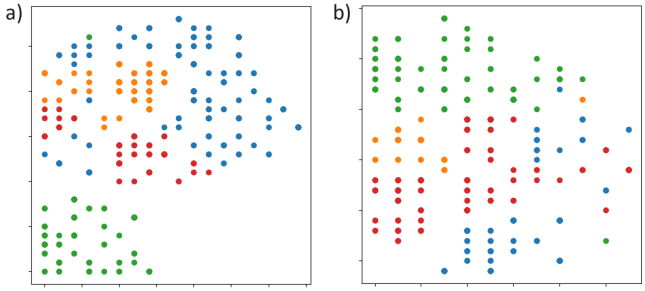

Figure 4: SOM clustering in a) CDSS dataset; b) CNLG dataset

Experiment results

To get initial information on predicted embeddings of users we have employed the use of TSNE visualization (see Fig. 3). Here, the individual colours represent different users. In either dataset, an interesting observation was made where the cases are grouped in two different types of clusters. We observed that the two clusters are relative to the type of users (i.e. single users or multiple users). While the cluster representing single-users showcases unique behaviour (possibly biased), the multi-user clusters may be considered relatively "stable" cognitive states. Next, SOM-based clustering (see Fig. 4 for 4 clusters) shows rather good identification of consistent areas in SOM space with stable borders and appearance of space between representing areas (see, e.g. "green" cluster in Fig. 4). Finally, running the evaluation for different numbers of clusters (see Fig. 5 and Table 1) shows that the proposed approach provides a significant decrease in

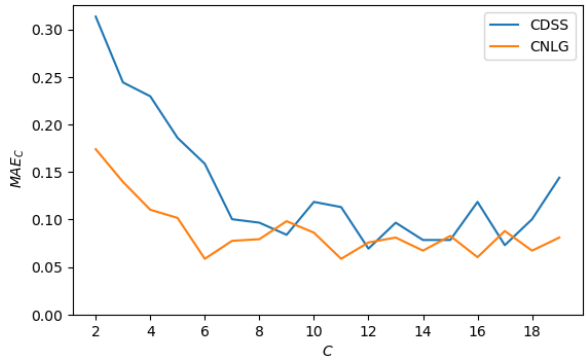

Figure 5: MAE_C dynamic depending on the number of clusters C in different datasets

$MAE_C(S_U)$ by 76-77% in comparison to $MAE_0(S_U)$ which is achieved with approximately $C \geq 12$ for CDSS and $C \geq 6$ for CNLG with no further significant decrease. This may be considered as an implicit evidence of different expertise diversity depending on problem domain.

Discussions and Future Work

The observed results indicate that we can predict the interconnected evaluation within the perception state model with relatively good performance which may be further improved by including information about the personal characteristics of decision-makers. The proposed approach can be used to better understand the internal reasoning and cognitive states of a decision-maker during HAI and human-AI collaboration. We believe there are several important aspects yet to be considered within such approaches.

First, the considered cognitive state is presumed to be more complicated than just a single latent variable. Existing approaches like the theory of mind, cognitive architectures, and inverse reinforcement learning can be employed to extend the model structure. Moreover, implicit characteristics may be indirectly identified through explicit personal characteristics (i.e. skills, position, certification, survey, etc.). We believe that extracting and correlating such ideas of describing cognitive states as a hierarchy of elements can prove to be more effective in modelling the actual cognitive state during more complex decision making scenarios (Jeung & Huang, 2023).

In our data implementation, the technology acceptance model was the most aligned and suitable model that can introduce the user's state during HAI in a simple yet sufficiently knowledgeable format. We believe that the user's cognitive state can be expressed in terms of descriptive and subjective scales which can be thereby interpreted as linear measurements to improve the overall dimensional interaction with the AI model. However, we note that certain scales of metrics might not be universally relevant to define a cognitive state, hence the usage of a hierarchy or sub-model structure can allow for thresholding of cognitive characteristics with low significance respective to its usage.

Second, the evaluation procedure significantly depends on a particular case and working environment. In the considered cases, the datasets and environment for human evaluation were artificially curated and filtered. Real-life cases can be much more complicated and consist of multiple parameters to be taken into account. For example, the considered characteristics of AI perceiving the human user state may be affected by the complexity of the case, i.e. constraints for decision-making, social interaction, and various other aspects (Papenmeier, Hienert, Kammerer, Seifert, & Kern, 2023). Thus, the external setting of such a model may have a significant influence on its internal structure.

Third, an important goal of the proposed approach is the structuring and portrayal of the human user's perceptional state which can be used for AI behavior strategy selection (e.g. implemented in the form of reinforcement learning to understand the reasoning, justification and thought process of the human user). Nevertheless, to provide a tool for collaboration, the values and goals of human users during decision making must be accounted for.

Further, an important aspect is the connection of internal evaluation of AI results (e.g. assessment of uncertainty, confidence, explainability, etc.) with predicted human perception. This aspect is key to estimating the level of complexity and explainability of the information provided by AI agents during human-AI collaboration. This is specifically crucial in complex domains to keep a personalized balance between the information required for proper decision-making and possible redundant information (Bach, Nørgaard, Brok, & van Berkel, 2023; Calisto et al., 2023).

Lastly, it is crucial to take into account different aspects of trust-based AI development concepts such as general AI trust, trust-to-particular-AI agent, and selective claims introduced by AI in certain cases. These levels of trust could also be considered as an influence on the human's perceptional state and be subjected to optimization.

Conclusion and future work

The presented work shows early results in the investigation of cognitive state roles in human-centric evaluation of AI recommendations. With the obtained preliminary results, we observed that the proposed approach to model subjective metrics may be used to attain an enhanced understanding of the perceptional process in humans when explicitly analysing cognitive states. This approach could be crucial and effective when utilised in scenarios that require significant expert insight or complex domains that require a multitude of expert experiences. In such cases, the proposed approach may be used for modelling human behaviour and their perceptional state during HAI and furthermore to introduce more human-centred metrics aimed at improved offline AI training.

The work described in this paper is an ongoing study aimed at developing approaches for improving HAI. We plan to continue the research in the direction of the issues mentioned above. Additionally, we aim to extend our study to additional domains and case studies.

Acknowledgements

The research was supported by by The Russian Science Foundation, Agreement №24-11-00272, https://rscf.ru/project/24-11-00272/.

References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... Horvitz, E. (2019). Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–13). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3290605.3300233 doi: 10.1145/3290605.3300233

Bach, A. K. P., Nørgaard, T. M., Brok, J. C., & van Berkel, N. (2023). "if i had all the time in the world": Ophthalmologists' perceptions of anchoring bias mitigation in clinical ai support. In *Proceedings of the 2023 chi conference on human factors in computing systems.* New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3544548.3581513 doi: 10.1145/3544548.3581513

Breck, E. J., Burger, J. D., Ferro, L., Hirschman, L., House, D., Light, M., & Mani, I. (2000, May). How to evaluate your question answering system every day ... and still get real work done. In *Proceedings of the second international conference on language resources and evaluation (LREC'00).* Athens, Greece: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L00-1153/

Calisto, F. M., Fernandes, J. a., Morais, M., Santiago, C., Abrantes, J. a. M., Nunes, N., & Nascimento, J. C. (2023). Assertiveness-based agent communication for a personalized medicine on medical imaging diagnosis. In *Proceedings of the 2023 chi conference on human factors in computing systems.* New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3544548.3580682 doi: 10.1145/3544548.3580682

Dillon, A. (2006, March). Human acceptance of information technology. In *International encyclopedia of ergonomics and human factors - 3 volume set* (pp. 1153–1156). CRC Press. Retrieved from http://dx.doi.org/10.1201/9780849375477-241 doi: 10.1201/9780849375477-241

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. Retrieved from https://doi.org/10.1177/1745691612460685 (PMID: 26172965) doi: 10.1177/1745691612460685

Evtikhiev, M., Bogomolov, E., Sokolov, Y., & Bryksin, T. (2023). Out of the bleu: How should we assess quality of the code generation models? *Journal of Systems and Software*, *203*, 111741. Retrieved from https://www.sciencedirect.com/science/article/pii/S016412122300136X doi: https://doi.org/10.1016/j.jss.2023.111741

Howes, A., Jokinen, J. P. P., & Oulasvirta, A. (2023, September). Towards machines that understand people. *AI Magazine*, *44*(3), 312–327. Retrieved from https://doi.org/10.1002/aaai.12116 doi: 10.1002/aaai.12116

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110. Retrieved from https://doi.org/10.1016/j.cobeha.2019.04.010 (Artificial Intelligence) doi: 10.1016/j.cobeha.2019.04.010

Jeung, J. L., & Huang, J. Y.-C. (2023). Correct me if i am wrong: Exploring how ai outputs affect user perception and trust. In *Companion publication of the 2023 conference on computer supported cooperative work and social computing* (pp. 323–327). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3584931.3606997 doi: 10.1145/3584931.3606997

Kazerooni, H. (1993). Extender: a case study for human-robot interaction via transfer of power and information signals. In *Proceedings of 1993 2nd ieee international workshop on robot and human communication* (pp. 10–20). Retrieved from https://ieeexplore.ieee.org/abstract/document/367756 doi: 10.1109/ROMAN.1993.367756

Kotseruba, I., & Tsotsos, J. K. (2018, July). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*(1), 17–94. Retrieved from http://dx.doi.org/10.1007/s10462-018-9646-y doi: 10.1007/s10462-018-9646-y

Kovalchuk, S. V., Kopanitsa, G. D., Derevitskii, I. V., Matveev, G. A., & Savitskaya, D. A. (2022). Three-stage intelligent support of clinical decision making for higher trust, validity, and explainability. *Journal of Biomedical Informatics*, *127*, 104013. Retrieved from https://www.sciencedirect.com/science/article/pii/S1532046422000296 doi: https://doi.org/10.1016/j.jbi.2022.104013

Kovalchuk, S. V., Lomshakov, V., & Aliev, A. (2022, December). Human perceiving behavior modeling in evaluation of code generation models. In A. Bosselut et al. (Eds.), *Proceedings of the 2nd workshop on natural language generation, evaluation, and metrics (gem)* (pp. 287–294). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.gem-1.24 doi: 10.18653/v1/2022.gem-1.24

Maclure, J. (2021, August). AI, explainability and public reason: The argument from the limitations of the human mind. *Minds and Machines*, *31*(3), 421–438. Retrieved from https://doi.org/10.1007/s11023-021-09570-x doi: 10.1007/s11023-021-09570-x

Miller, G. A. (1956, March). The magical number seven, plus or minus two: Some limits on our capacity for

processing information. *Psychological Review*, *63*(2), 81–97. Retrieved from http://dx.doi.org/10.1037/h0043158 doi: 10.1037/h0043158

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., ... Schulman, J. (2021). *Webgpt: Browser-assisted question-answering with human feedback.* arXiv. Retrieved from https://arxiv.org/abs/2112.09332 doi: 10.48550/ARXIV.2112.09332

Papenmeier, A., Hienert, D., Kammerer, Y., Seifert, C., & Kern, D. (2023). Know what not to know: Users' perception of abstaining classifiers. In *Companion publication of the 2023 acm designing interactive systems conference* (pp. 169–172). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3563703.3596622 doi: 10.1145/3563703.3596622

Sreedharan, S., Kulkarni, A., Smith, D., & Kambhampati, S. (2021, 8). A unifying bayesian formulation of measures of interpretability in human-ai interaction. In Z.-H. Zhou (Ed.), *Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21* (pp. 4602–4610). International Joint Conferences on Artificial Intelligence Organization. Retrieved from https://doi.org/10.24963/ijcai.2021/625 (Survey Track) doi: 10.24963/ijcai.2021/625

Sundar, S., Kim, J., & Gambino, A. (2017). Using theory of interactive media effects (time) to analyze digital advertising. In *Digital advertising* (pp. 86–109). United States: Taylor and Francis. Retrieved from https://www.taylorfrancis.com/chapters/edit/10.4324/9781315623252-6/ doi: 10.4324/9781315623252-6

Taha, A. A., & Hanbury, A. (2015, August). Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, *15*(1). Retrieved from https://doi.org/10.1186/s12880-015-0068-x doi: 10.1186/s12880-015-0068-x