Effect of Abstraction on Understandability of ASP Explanations: Work in Progress

Zeynep G. Saribatur[1], Johannes Langer[2] and Ute Schmid[2]

[1] Institute of Logic and Computation, TU Wien
[2] Cognitive Systems, University of Bamberg

Abstract. We report on our ongoing work on investigating the effect of abstraction on the understandability of ASP explanations, by considering the recent abstraction notions over irrelevancy. We describe our hypotheses and cognitive experiment design. Our preliminary study showed no significant effect resulting in further investigations. A recent extension of the study, that overcomes the limitations of the previous study, now shows significant effect of abstraction, while some open questions remain.

1 Abstractions over Explanations

Obtaining explanations of answer sets has been a topic that is widely studied with available systems. The challenge of achieving concise explanations to aid in human understanding still remains [3]. Recent works in ASP abstraction study the theory of removal [6] and clustering [4] of irrelevant details in answer set programs. A simplification, resp., an abstraction, of an answer set program is defined that preserves the dependencies according to a uniform equivalence-like notion while reducing the vocabulary by removal, resp., clustering.

Definition 1 ([4]) *Given sets of atoms $\mathcal{U}, \mathcal{U}'$ with $|\mathcal{U}| \geq |\mathcal{U}'|$, a program P over $\mathcal{U}$ and a mapping $m : \mathcal{U} \mapsto \mathcal{U}'$, Q over $\mathcal{U}'$ is a* uniform m-abstraction of P *if, for any set F of facts over $\mathcal{U}$, we have*

$$m(AS(P \cup F)) = AS(Q \cup m(F)). \tag{1}$$

Informally, the aim is to map atoms from the language $\mathcal{U}$ of program P to atoms in a smaller language $\mathcal{U}'$ (removal maps atoms to $\top$, while clustering maps multiple atoms to a clustered atom) in such a way that the answer sets of P and the resulting abstraction Q correspond, independently of the facts, i.e., the instance data, added to the program.

We hypothesize that the explanations obtained in the abstract programs would help in understandability, due to only containing the relevant details compared to the default explanations of the original programs. Next, we describe the cognitive experiment in form of an online survey which we designed to test our hypothesis and report on preliminary results[3] and the recent results of the updated experiment addressing the previous limitations.

[3] These results appeared as a Technical Communication at KI 2025 [5].

2 Empirical Study

Participants are presented with a classification task of tabular data, where each instance is a set of attributes, also visualized by a corresponding image. We present three concept learning tasks to participants, where each task is defined through an answer set program, assigning the instance to the target class when certain conditions are met. The presented explanations on classifications are obtained from `xclingo` [2]. During evaluation, participants need to decide for each instance whether it belongs to the target class or not. We have three domains: flowers, mushrooms, and cacti, which six domain-specific decision attributes, and the same target attribute "dangerous".

The empirical online study is based on a complete 2x2 between-subject factorial design, where participants were randomly assigned to one of the four groups formed by a combination of 2 factors: 'cluster' and 'removal', referring to the abstraction of the answer set programs w.r.t clustering [4] and removal [6], and the data between these groups were compared to calculate the effect of each factor on our dependent variable. The main part of the experiment consists of a learning and a test phase. We recruited 71 participants from two universities and tested the hypothesis "Having abstract explanations during the learning phase increases the classification accuracy compared to the `control` group", though statistical tests showed no significant effect.

We detected at least two unforeseen limitations of this study: The semantic meaning of the target variable "dangerous" may have resulted in participants choosing dangerousness to avoid repercussions, which was also reported among the comments. Furthermore, the shown images might have been more helpful than expected, helping humans abstract over details, thus not necessarily needing the abstracted explanations.

Updated Study We updated the study design by changing the target variable to less triggering terms for each domain and removing the figures. Furthermore we added comprehension and attention checks, making sure to have participants that understand the task and pay attention. We made the survey available on Prolific, and 100 participants have completed the tasks. The statistical tests show significant effect of clustering in performance, while there is significant effect of removal on response times (RT). Overall when considering abstraction as a one-factor, we also observe significant effect in performance and in RT.

3 Discussion

Our results show that abstraction by clustering helps in understanding, while abstraction by removal reduces cognitive effort. This dual effect of abstraction types raises questions on the cognitive meaning of these abstraction notions and whether the complexity of the explanations has a role [1], which will need to be explored. Our investigations also show the challenge of capturing the effect of abstraction on human understandability, as the experiment design might cause unexpected results due to unforeseen perceptions or abilities of humans.

References

1. Ai, L., Muggleton, S.H., Hocquette, C., Gromowski, M., Schmid, U.: Beneficial and harmful explanatory machine learning. Machine Learning **110**, 695–721 (2021)
2. Cabalar, P., Muñiz, B.: Explanation graphs for stable models of labelled logic programs. In: Arias, J., Batsakis, S., Faber, W., Gupta, G., Pacenza, F., Papadakis, E., Robaldo, L., Rückschloß, K., Salazar, E., Saribatur, Z.G., Tachmazidis, I., Weitkämper, F., Wyner, A.Z. (eds.) Proceedings of the International Conference on Logic Programming 2023 Workshops co-located with the 39th International Conference on Logic Programming (ICLP 2023), London, United Kingdom, July 9th and 10th, 2023. CEUR Workshop Proceedings, vol. 3437. CEUR-WS.org (2023), https://ceur-ws.org/Vol-3437/paper3ASPOCP.pdf
3. Fandinno, J., Schulz, C.: Answering the "why" in answer set programming - A survey of explanation approaches. TPLP **19**(2), 114–203 (2019)
4. Saribatur, Z.G., Knorr, M., Gonçalves, R., Leite, J.: On abstracting over the irrelevant in answer set programming. In: Marquis, P., Ortiz, M., Pagnucco, M. (eds.) Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024 (2024). https://doi.org/10.24963/KR.2024/61, https://doi.org/10.24963/kr.2024/61
5. Saribatur, Z.G., Langer, J., Thaler, A.M., Schmid, U.: Towards observing the effect of abstraction on understandability of explanations in answer set programming. In: Braun, T., Paaßen, B., Stolzenburg, F. (eds.) KI 2025: Advances in Artificial Intelligence. pp. 236–243. Springer Nature Switzerland, Cham (2026)
6. Saribatur, Z.G., Woltran, S.: Foundations for Projecting Away the Irrelevant in ASP Programs. In: Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning. pp. 614–624. IJCAI Organization (2023). https://doi.org/10.24963/kr.2023/60