In the Hand of the Beholder: Comparing Interactive Proof Visualizations (Extended Abstract)*

Christian Alrabbaa[1], Stefan Borgwardt[1], Nina Knieriemen[2], Alisa Kovtunova[1], Anna Milena Rothermel[2], and Frederik Wiehr[2]

[1] Institute of Theoretical Computer Science, TU Dresden, Germany
firstname.lastname@tu-dresden.de
[2] German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus, Saarbrücken, Germany
firstname[_middlename].lastname@dfki.de

1 Introduction

Although logical inferences are interpretable, actually explaining them to a user is still a challenging task. While sometimes it may be enough to point out the axioms from the ontology that lead to the consequence of interest, more complex inferences require proofs with intermediate steps that the user can follow [19]. Following a line of research on the understandability of description logic inferences and proofs [2–4, 9, 12, 14, 15, 18], in this paper we investigate the usefulness of different proof representations. We compare proofs in a traditional tree shape with a linearized textual representation of proofs. For both variants, we provide an interactive as well as a static version.

Our main goal in this paper is to find out whether a user's cognitive ability influences which of these four proof representations are preferred and lead to better performance. A study from last year [9] attempted a similar comparison, but it did not find significant differences based on the user's self-reported experience with logic. In this paper, we attempt to measure the user's cognitive ability level and investigate the impact on performance and preferences of different proof representations. We report here about two experiments. For both we were able to increase (cf. [9]) the number of participants by creating a fully online survey on LimeSurvey[3] and making it available at a study participant recruitment platform called Prolific[4].

In a first experiment, we verified that the score on the standardized 16-item International Cognitive Ability Resource (ICAR16) test[5] strongly correlates with the ability to draw logical inferences and understand logical proofs. Based on this insight, we used the ICAR16 in our second experiment to measure the user's cognitive ability levels and compare the different proof representations.

* This is an abstract of the paper [6] accepted at DL 2021.
[3] https://www.limesurvey.org/
[4] https://www.prolific.co/
[5] https://icar-project.com/projects/icar-project/wiki

2 Proofs

We assume a basic familiarity with Description Logics (DLs), in particular $\mathcal{ALCQ}$ [8]. Let $\mathcal{O}$ be an ontology and α a consequence of $\mathcal{O}$ ($\mathcal{O} \models \alpha$). The next step is to compute *justifications*, i.e., minimal subsets $\mathcal{J} \subseteq \mathcal{O}$ such that $\mathcal{J} \models \alpha$, which already point out the axioms from $\mathcal{O}$ that are responsible for α. However, actually understanding why α follows may require a more detailed proof, see [2,3] for the formal proof framework. It is important that proofs are neither too detailed nor too short [9,19,20].

Concerning the representation of logical statements, it has been observed by [17] that statements in a controlled natural language are understood significantly better than the Manchester OWL Syntax, where DL axioms are expressed by sentences with the words like "SubTypeOf", "DisjointWith", "HasDomain", etc. Therefore, to open our experiments to a larger population, similarly to the approaches [1,17,20] we use patterns to convert DL sentences into natural-language explanations. Moreover, we use nonsense names that vaguely look and sound English to enable more natural-sounding sentences, e.g. "Every woal is munted only with luxis that are kakes". We did not use real words, because we already faced a problem concerning prior knowledge about the example domains in [9].

Further, we (1) arrange these sentences in a tree-shaped representation, similarly to proofs based on consequence-based reasoning procedures [16, 21], and we (2) order them in a linear sequence using English conjunctions, e.g. as produced by verbalization techniques [7,17,18,20]. An aspect in which text differs from a proof tree are that conjunctions (e.g. "since", "and") are used to illustrate proof steps and that statements may be repeated if they are reused later.

The interactive proofs were provided by a prototypical web application for explaining DL entailments called Evonne [5, 13]. There are examples of the interactive text[6] and the interactive tree representation[7] we used online.

3 Connecting Logical Abilities and Proof Understanding

We first conducted an experiment that shows a connection between participants' understanding of logical proofs and their general cognitive abilities. A printable version of the survey is available online.[8]

3.1 Description of the experiment

For this experiment we want to employ a standardized measure that allows us to predict the understanding of logical proofs. Here our *hypothesis* is that the ICAR16 score predicts the performance in the logical reasoning tasks.

[6] `https://lat.inf.tu-dresden.de/evonne/textProof1`
[7] `https://lat.inf.tu-dresden.de/evonne/proof1`
[8] `https://cloud.perspicuous-computing.science/s/oHp9pRaoCx5SDsF`

Participants. The sample consisted of 101 participants (45 female, 56 male) with a mean age of $M = 24.52$ ($SD = 6.81$).

Material. To assess the participants' cognitive abilities, the 16-item International Cognitive Ability Resource (ICAR16) [11, 22] was applied. It consists of 16 questions equally distributed over four different types: matrix reasoning, letter and number series, verbal reasoning, and 3-dimensional rotation.[5] The maximum score was 1, while the minimum score was 0.

To test the performance with formal proofs, participants had to solve two tasks. The first described a set of 16 axioms (in natural language) and they should decide which of the given statements follow from the axioms. Each of the statements could be marked as "follows", "does not follow" or "I do not know". In the second task, they were given a proof in tree shape that contained a blank node, and they were asked which of 8 given statements would be valid labels for the node in the context of the proof ("yes", "no", "I do not know"). The highest possible score a participant can achieve was 24.

3.2 Results

The mean of the ICAR16 scores was $M = .55$ ($SD = .24$). The mean of the score for both logical reasoning tasks was $M = 15.99$ ($SD = 3.3$).

A multiple regression analysis was carried out using the performance in the logical reasoning tasks as the dependent and the ICAR16 performance as the independent variable. The ICAR16 score significantly predicted the performance in the logical tasks ($F(1, 99) = 43.15$, $p < .001$). The ICAR16 explained 30% of the variation in the score of the logical tasks ($R^2 = .3$, $p < .001$), which can be interpreted as large effect size/high explained variance [10]. In other words, there is evidence that cognitive abilities determine the performance in logical tasks.

4 Logical Abilities and Proof Representation Preferences

Given that ICAR16 scores are highly correlated with performance on logical reasoning tasks, we used it in our main experiment to distinguish participants by their cognitive ability level. A printable version of the survey is available online.[9]

4.1 Description of the experiment

With this experiment, we attempt to find out which proof representation is most understandable for different users. The goal is to find a difference in the (subjective) preferences and (objective) performance on each proof representation, depending on the user's level of cognitive ability.

Hypothesis 1: It is easier to understand interactive proofs than static proofs. This will be shown by an increase in performance and by a higher comprehensibility rating for the interactive conditions.

[9] `https://cloud.perspicuous-computing.science/s/dCSmbraoJ4RzDqG`

Hypothesis 2: The relative level of comprehensibility of a tree-shaped vs. textual proof depends on the cognitive abilities. This will be shown by a difference in performance and difficulty rating between the conditions and in the final ranking, in dependence of the ICAR16 scores.

Participants. The final sample consisted of 173 participants (41% female, 59% male) with a mean age of $M = 24.8$ ($SD = 8.21$).

Material. Again, we used ICAR16 to assess the participants' cognitive abilities.

For the proof representations, there were two different conditions: form (tree-shaped or textual) and interactivity (static or interactive). We used a 2×2 within-subjects design, which means that each participant saw all four condition combinations. For this, we developed four artificial proofs of roughly the same difficulty level. For each proof, there were three pages of questions. Each question page contained a single question with 6 answer options (plus "none of these" and "I don't know"). Questions were of the form "Which of the following would be a correct replacement for the deduction 'XYZ' in the proof?" or "Which parts of the following summary/reformulation of the proof are incorrect?" In the end, a score was calculated based on the number of correct answers (out of 12).

In addition to the performance tasks, participants had to rate each proof according to its comprehensibility on a scale from 1 ("not at all") to 5 ("very much"). In the end of the survey, they would also rank the comprehensibility of the four different proof representations compared to each other.

4.2 Results

A median split ($mdn = .44$) was carried out to divide the participants into those who achieved high scores in the ICAR16 and thus presumably also have higher cognitive abilities and those who scored lower.

Performance and Comprehensibility Ratings. To compare the performance and the comprehensibility ratings after each proof, we ran a multivariate analysis of variance (MANOVA). We could not detect differences in the comprehensibility ratings as well as in the performance between the various representations. Neither of our two hypotheses could be conclusively confirmed.

Ranking. To evaluate the ranking of the four representations (1 = most comprehensible, 4 = least comprehensible), we ran a Friedman's test revealing a significant difference across both ICAR16 groups. Post-hoc pairwise comparisons were Bonferroni-corrected and showed three significant comparisons. The interactive tree was significantly more often ranked higher than the interactive text ($z = .40$, $p = .024$, Cohen's effect size $r = .03$) and also higher than static text ($z = -.50$, $p = .002$, Cohen's effect size $r = .04$). The static tree representation was also ranked significantly higher than static text, $z = .39$, $p = .032$, Cohen's effect size $r = .03$ (see Figure 1, the light part).

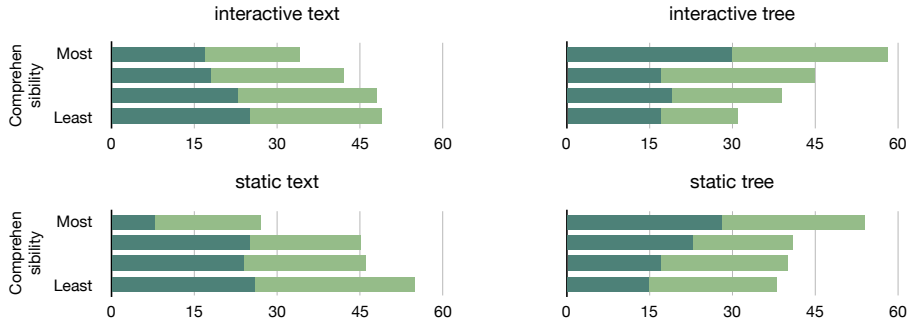

Fig. 1. Rankings of all participants (full length of green bars) and of those with higher ICAR scores (dark parts) for each condition combination.

A Friedman's test in the group with higher ICAR performance showed a significant difference in the ranking of representations. Bonferroni-corrected post-hoc pairwise comparisons revealed two significant comparisons: between static tree and static text ($z = .59$, $p = .019$, Cohen's effect size $r = .06$) with static tree being ranked higher than static text (see Figure 1, the dark bars). Interactive tree was also preferred before static text, ($z = -.54$, $p = .041$, Cohen's effect size $r = .06$). In contrast, the low-ICAR-performers showed no significant difference.

Limitations. According to the aims of our study, we did not pre-select participants. 55.5% of the participants had no experience with propositional logic. For many participants the proof tasks were very challenging, resulting in a mean score of $M = 2.36$ (out of 12). This resulted in many data points being clustered on the lower end of the scale and differences being more difficult to detect.

5 Conclusion

In addition to previous observations that shorter proofs are better [9, 20], we observed a subjective preference for tree-shaped proofs, although this was not reflected by increased performance in our study. Moreover, the level of cognitive abilities did not seem to influence the preferences or the subjective ratings. As a side result, we demonstrated that cognitive abilities tested by the ICAR16 predict the reasoning performance in formal logics. In future work, we want to further investigate the trade-off between giving no details (i.e., justifications) and giving too many details (i.e., full proofs) in various representation formats. For laypersons, it may be better to quickly communicate the gist of a proof in natural language, whereas experts may require access to the formal details.

Acknowledgements This work was partially supported by DFG grant 389792660 as part of TRR 248 (`https://perspicuous-computing.science`), and QuantLA, GRK 1763 (`https://lat.inf.tu-dresden.de/quantla`).

References

1. Alharbi, E., Howse, J., Stapleton, G., Hamie, A., Touloumis, A.: The efficacy of OWL and DL on user understanding of axioms and their entailments. In: d'Amato, C., Fernández, M., Tamma, V.A.M., Lécué, F., Cudré-Mauroux, P., Sequeda, J.F., Lange, C., Heflin, J. (eds.) The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part I. Lecture Notes in Computer Science, vol. 10587, pp. 20–36. Springer (2017). https://doi.org/10.1007/978-3-319-68288-4_2
2. Alrabbaa, C., Baader, F., Borgwardt, S., Koopmann, P., Kovtunova, A.: Finding small proofs for description logic entailments: Theory and practice. In: Albert, E., Kovacs, L. (eds.) LPAR-23: 23rd International Conference on Logic for Programming, Artificial Intelligence and Reasoning. EPiC Series in Computing, vol. 73, pp. 32–67. EasyChair (2020). https://doi.org/10.29007/nhpp
3. Alrabbaa, C., Baader, F., Borgwardt, S., Koopmann, P., Kovtunova, A.: On the complexity of finding good proofs for description logic entailments. In: Borgwardt, S., Meyer, T. (eds.) Proceedings of the 33rd International Workshop on Description Logics (DL 2020) co-located with the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR 2020), Online Event [Rhodes, Greece], September 12th to 14th, 2020. CEUR Workshop Proceedings, vol. 2663. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2663/paper-1.pdf
4. Alrabbaa, C., Baader, F., Borgwardt, S., Koopmann, P., Kovtunova, A.: Finding good proofs for description logic entailments using recursive quality measures (extended technical report). CoRR **abs/2104.13138** (2021), https://arxiv.org/abs/2104.13138
5. Alrabbaa, C., Baader, F., Dachselt, R., Flemisch, T., Koopmann, P.: Visualising proofs and the modular structure of ontologies to support ontology repair. In: Borgwardt, S., Meyer, T. (eds.) Proceedings of the 33rd International Workshop on Description Logics (DL 2020) co-located with the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR 2020), Online Event [Rhodes, Greece], September 12th to 14th, 2020. CEUR Workshop Proceedings, vol. 2663. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2663/paper-2.pdf
6. Alrabbaa, C., Borgwardt, S., Knieriemen, N., Kovtunova, A., Rothermel, A.M., Wiehr, F.: In the hand of the beholder: Comparing interactive proof visualizations (2021), accepted at the 34th International Workshop on Description Logics (DL 2021)
7. Androutsopoulos, I., Lampouras, G., Galanis, D.: Generating natural language descriptions from OWL ontologies: The NaturalOWL system. Journal of Artificial Intelligence Research **48**, 671–715 (2013). https://doi.org/10.1613/jair.4017
8. Baader, F., Horrocks, I., Lutz, C., Sattler, U.: An Introduction to Description Logic. Cambridge University Press (2017). https://doi.org/10.1017/9781139025355
9. Borgwardt, S., Hirsch, A., Kovtunova, A., Wiehr, F.: In the Eye of the Beholder: Which Proofs are Best? In: Borgwardt, S., Meyer, T. (eds.) Proc. of the 33rd Int. Workshop on Description Logics (DL 2020). CEUR Workshop Proceedings, vol. 2663 (2020), http://ceur-ws.org/Vol-2663/paper-6.pdf
10. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, 2nd edn. (1988). https://doi.org/10.4324/9780203771587
11. Condon, D.M., Revelle, W.: The international cognitive ability resource: Development and initial validation of a public-domain measure. Intelligence **43**, 52–64 (2014). https://doi.org/10.1016/j.intell.2014.01.004

12. Engström, F., Nizamani, A.R., Strannegård, C.: Generating comprehensible explanations in description logic. In: Informal Proceedings of the 27th International Workshop on Description Logics, Vienna, Austria, July 17-20, 2014. pp. 530–542 (2014), http://ceur-ws.org/Vol-1193/paper_17.pdf
13. Flemisch, T., Langner, R., Alrabbaa, C., Dachselt, R.: Towards designing a tool for understanding proofs in ontologies through combined node-link diagrams. In: Ivanova, V., Lambrix, P., Pesquita, C., Wiens, V. (eds.) Proceedings of the Fifth International Workshop on Visualization and Interaction for Ontologies and Linked Data co-located with the 19th International Semantic Web Conference (ISWC 2020), Virtual Conference (originally planned in Athens, Greece), November 02, 2020. CEUR Workshop Proceedings, vol. 2778, pp. 28–40. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2778/paper3.pdf
14. Horridge, M., Parsia, B., Sattler, U.: Justification oriented proofs in OWL. In: The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I. pp. 354–369 (2010). https://doi.org/10.1007/978-3-642-17746-0_23
15. Kazakov, Y., Klinov, P., Stupnikov, A.: Towards reusable explanation services in Protege. In: Artale, A., Glimm, B., Kontchakov, R. (eds.) Proc. of the 30th Int. Workshop on Description Logics (DL'17). CEUR Workshop Proceedings, vol. 1879 (2017), http://www.ceur-ws.org/Vol-1879/paper31.pdf
16. Kazakov, Y., Krötzsch, M., Simancik, F.: The incredible ELK – from polynomial procedures to efficient reasoning with $\mathcal{EL}$ ontologies. J. Autom. Reasoning **53**(1), 1–61 (2014). https://doi.org/10.1007/s10817-013-9296-3
17. Kuhn, T.: The understandability of OWL statements in controlled english. Semantic Web **4**(1), 101–115 (2013). https://doi.org/10.3233/SW-2012-0063
18. Nguyen, T.A.T., Power, R., Piwek, P., Williams, S.: Measuring the understandability of deduction rules for OWL. In: Proceedings of the First International Workshop on Debugging Ontologies and Ontology Mappings, WoDOOM 2012, Galway, Ireland, October 8, 2012. pp. 1–12 (2012), http://www.ida.liu.se/~patla/conferences/WoDOOM12/papers/paper4.pdf
19. Schiller, M.R.G., Glimm, B.: Towards explicative inference for OWL. In: Informal Proceedings of the 26th International Workshop on Description Logics, Ulm, Germany, July 23 - 26, 2013. pp. 930–941 (2013), http://ceur-ws.org/Vol-1014/paper_36.pdf
20. Schiller, M.R.G., Schiller, F., Glimm, B.: Testing the adequacy of automated explanations of EL subsumptions. In: Proceedings of the 30th International Workshop on Description Logics, Montpellier, France, July 18-21, 2017. (2017), http://ceur-ws.org/Vol-1879/paper43.pdf
21. Simancik, F., Kazakov, Y., Horrocks, I.: Consequence-based reasoning beyond horn ontologies. In: IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011. pp. 1093–1098 (2011). https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-187
22. Young, S.R., Keith, T.Z.: An examination of the convergent validity of the ICAR16 and WAIS-IV. Journal of Psychoeducational Assessment **38**(8), 1052–1059 (2020). https://doi.org/10.1177/0734282920943455