




In the Head of the Beholder: Comparing Different Proof Representations

Christian Alrabbaa¹, Stefan Borgwardt¹, Anke Hirsch², Nina Knieriemen²,
Alisa Kovtunova¹, Anna Milena Rothermel², and Frederik Wiehr²

¹ Institute of Theoretical Computer Science, TU Dresden, Germany
`firstname.lastname@tu-dresden.de`

² German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
`firstname[_middlename].lastname@dfki.de`

Abstract. Ontologies provide the logical underpinning for the Semantic Web, but their consequences can sometimes be surprising and must be explained to users. A promising kind of explanations are proofs generated via automated reasoning. We report about a series of studies with the purpose of exploring how to explain such formal logical proofs to humans. We compare different representations, such as tree- vs. text-based visualizations, but also vary other parameters such as length, interactivity, and the shape of formulas. We did not find evidence to support our main hypothesis that different user groups can understand different proof representations better. Nevertheless, when participants directly compared proof representations, their subjective rankings showed some tendencies such as that most people prefer short tree-shaped proofs. However, this did not impact the user’s understanding of the proofs as measured by an objective performance measure.

Introduction

Explanations of automated decisions are currently an important topic of research. However, apart from the discussion about how explainable different AI methods are, the main task of explanations is *understanding*, i.e. that the information transmitted is actually received by the human user [32]. Even methods that are “explainable by design”, such as logic-based ones, are not necessarily understandable by design when presenting them to laypersons.

In the area of Description Logics (DLs) [10], research on explanations first focused on proofs for explaining logical consequences [13, 30], but it was quickly realized that often it is enough to point out a minimal set of responsible axioms from the ontology, i.e. so-called *justifications* [11, 21, 37]. While justifications are already very helpful for designing or debugging an ontology, depending on the complexity of the inference and the expertise of the user, more detailed proofs are needed to fully understand why the consequence follows from the axioms. Therefore, researchers have thought about providing (partial) proofs [23, 26] and developed more user-friendly presentation formats, e.g. using natural language instead of logical formulas [33–35].

Following a line of research on the understandability of description logic inferences and proofs [3–5,18,23,26,33,34], in this paper we compare the usefulness of different proof representations. In an effort to understand which approaches are most promising for improving explainability, we studied which representations of DL proofs are preferred by users (with and without prior experience in logic) and which of them actually lead to an increased performance when doing logic-related tasks. In this paper, we summarise the lessons learned after conducting four experiments. All studies use proofs in a traditional tree shape, e.g. based on consequence-based reasoning procedures [27,39], and linearized translations of these proofs into text, e.g. as done by various verbalization techniques [8,33,35]. These conditions are representative of the state-of-the-art in DL explanations. We hand-crafted all proofs for the studies, but tried to stay as close as possible to the actual output of these systems. The main goal throughout these studies was to find differences in user preferences between different user groups. Our *main hypothesis* was that users with a different level of experience with logic would work better with different proof representations, e.g. text- vs. tree-based ones. While this was not confirmed, we gained some insights about subjective preferences of proof presentations, e.g. that short, tree-shaped proofs are preferred in general.

Related Work. Several approaches for converting description logic axioms and proofs into textual representations have been developed and evaluated [1,8,29,34,35]. For example, generation of verbalized explanations for non-trivial derivations in a real world domain was tested on computer scientists in [35]. The authors distinguish short and long textual explanations, but the participants’ opinions on conciseness turned out to be mixed and not too strong. In [29], it has been confirmed that statements in a controlled natural language are understood significantly better than the *Manchester OWL Syntax*, where DL axioms are expressed by sentences with words like “SubTypeOf”, “DisjointWith”, “HasDomain”, etc. Moreover, the experiment [1] has shown that the Manchester syntax is not more effective than the formal DL syntax. Differently from previous studies [33–35], in most of our experiments we directly compared textual and tree proof formats. In [28], the authors look into various hybrid proof representations and evaluate them in terms of understanding. In contrast to our work, they focus on *defeasible logics*, they do not consider pure textual representations, and the user evaluation involved postgraduate students. The work described in [17] deals with explaining logical inconsistencies in a healthcare domain using natural language, but it does not consider graphical proof representations.

More details and printable versions of the surveys are available online.³ Studies I–III have previously been presented in workshop papers [7,14].

Background

The proofs we use are loosely based on the DL *ALCQ* [10], but deep knowledge of this logic is not required here. We denote DL statements (called *axioms*) by α

³ gitlab.perspicuous-computing.science/a.kovtunova/user-study-collection

$$\frac{\frac{A \sqsubseteq \exists r. \top \quad A \sqsubseteq \forall r. (B \sqcap C)}{A \sqsubseteq \exists r. (B \sqcap C)} \quad C \sqcap B \sqsubseteq \perp}{A \sqsubseteq \perp} \quad \mathcal{O} = \{ A \sqsubseteq \exists r. \top, \\ C \sqcap B \sqsubseteq \perp, \\ A \sqsubseteq \forall r. (B \sqcap C) \}$$

Fig. 1. A proof for the unsatisfiability of A w.r.t. \mathcal{O} , i.e. that $\mathcal{O} \models A \sqsubseteq \perp$.

Table 1. Different proof representations for our experiments.

Study	Text proofs	Length		Tree proofs			Domain		Letters
		Long	Short	DL syntax	Arrows	Real	Nonsense		
I	*	*	*	*		*			
II			*		*		*		
III	*		*		*		*		
IV	*		*		*			*	

and *ontologies*, which are finite sets of axioms, by \mathcal{O} . Let \mathcal{O} be an ontology and α be a consequence of \mathcal{O} (written $\mathcal{O} \models \alpha$). The first step towards understanding why this consequence holds is to compute *justifications* [11, 21, 37], 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. Informally, a *proof* is a tree consisting of inference steps $\frac{\alpha_1 \dots \alpha_n}{\alpha}$, where each step is sound, i.e. $\{\alpha_1, \dots, \alpha_n\} \models \alpha$ holds (see Figure 1). Often, such a proof is built from the *inference rules* of an appropriate calculus [9, 39]. However, there also exist approaches to generate DL proofs that start with a justification, and extend it with intermediate axioms (*lemmas*) using heuristics [21, 22], concept interpolation [36], or forgetting [3].

It is important that proofs are neither too detailed nor too short. In fact, a justification can itself be seen as a one-step proof of a consequence α , but if each element of the justifications seems reasonable to the user, then it can be hard to track down the precise interaction between these axioms that causes the problem. Axioms may not always behave as the user expects, e.g. “every A has only r s that are B s” ($A \sqsubseteq \forall r. B$) does not imply that “every A has an r that is a B ” ($A \sqsubseteq \exists r. B$). On the other hand, too many small proof steps can also be detrimental for understanding, because they are distracting. For example, it may happen that a reasoner includes the trivial step $\frac{C \sqcap B \sqsubseteq \perp}{B \sqcap C \sqsubseteq \perp}$ in Figure 1 to make the two conjunctions match syntactically, which may not be necessary for understanding the essence of the proof. Apart from proof length, in our experiments we also use other ways of varying the proof representations (see Table 1). For example, in Studies II–IV we use a more flexible visualization of trees in which arrows are used instead of horizontal lines (see the supplementary PDF file in the repository³).

A textual representation of a proof is necessarily a *linearization*, where the inference steps are explained in a sequence, for example in a top-down left-right order. A text corresponding to the tree proof in Figure 1 could be the following:

Table 2. Overview of the experiments

Study	1-on-1	Online ⁴	# participants			Avg. time (min)	Mean age (SD)	Pay
	interview	survey	male	female	non-binary			
I	*		12	4	–	90	23.0 (1.71)	20 €
II		*	56	45	–	29	24.5 (6.8)	£ 5.20
III		*	102	71	–	51	24.8 (8.2)	£ 8.75
IV		*	41	66	1	44	25.9 (6.9)	£ 6.25

Since every A has an r and every A has only rs that are Bs and Cs, every A has an r which is a B and a C. Since there is no object which is a C and a B at the same time, there is no object of type A.

Other aspects in which a 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.

We use the formal DL syntax for tree proofs only in the first experiment over axioms expressing medical knowledge, e.g. the statement “there is no object which is both a compound and an atom at the same time” is presented as the expression $\text{Atom} \sqcap \text{Compound} \sqsubseteq \perp$. In the later experiments, we do not use real domains to avoid interference from prior knowledge about the domain. We also adopt the approach from [29, 35] and avoid the formal syntax in order to include more participants. For example, in Study IV, $A \sqsubseteq \exists r. T$ would be shown as “Every A has an r.” In the remaining two experiments, 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” instead of “Every A has only rs that are Bs and Cs” ($A \sqsubseteq \forall r. (B \sqcap C)$); see also Table 1.

General Study Information. In Table 2 we summarize the demographic data for the experiments. All study participants were at least 18 years old. For the online surveys, we had to filter out participant answers of low quality. For this purpose, attention check questions, e.g. “In this statement, please choose ‘No.’” were introduced. To compute all quantitative analyses, IBM SPSS Statistics (Version 26) for Windows [24] and the Macro PROCESS [20] was used. For all hypotheses, we used a p -value threshold of 0.05.

Study I – Are Short Proofs Preferred?

We started our investigation of participants’ understanding of different proof representations by interviewing participants. Here, we used both textual proofs and classical tree proofs using DL syntax. To find out how detailed proofs should be, we used shortened versions for each of tree and text representations, in which some (easy) reasoning steps were omitted or merged. During the interviews, we observed whether participants’ understanding differs between these four condition

⁴ The participants were recruited using Prolific (<https://www.prolific.co/>). No restrictions on participant background were imposed.

combinations. Moreover, we wanted to investigate if experience in logic influences the performance and preferences.

Conditions and Design. We used two different conditions with two levels each. One condition was the representational form of the proof, which was either text or tree. The other condition was the length of the proof, which was either short or long. Thus, there were the four following condition combinations: *Long Text*, *Short Text*, *Long Tree*, and *Short Tree*. We used a 2×2 within-subjects design, which means that each participant saw all four representations on four different proofs following a Latin square design.³ The independent variable was the experience, while the dependent variable was the rating of the proofs.

Material. Proofs from the medical domain were chosen such that they represent unintuitive consequences, e.g. the unsatisfiability of a concept name, or that an amputation of a finger is also an amputation of the whole hand [12]. All four examples were chosen from the literature on DL explanations [12, 25, 31, 38]. For each of them, four different proof representations were manually created, not automatically generated, to make them comparable in difficulty.

To make sure the participants really understood the proofs, a logic expert reviewed the video of each participant after each session. We used the think-aloud technique, so the expert was able to follow the participant’s thoughts and rated the video based on the participant’s understanding on a scale from 1 (no understanding) to 3 (complete understanding).

Further Information. To assess participants’ experience, we asked them how they would rate their experience with propositional-, description-, and first-order logic on a scale from 1 (no knowledge) to 5 (expert). We evaluated how they rated the difficulty of each proof on a scale from 1 (very easy) to 5 (very difficult). To compare the proof representations, at the end of the experiment we asked the participants to rank the proofs based on their comprehensibility (first rank = very easy, fourth rank = very difficult). It was possible to give several proofs the same rank.

Participants (see Table 2). Our participants were recruited from undergraduate and graduate university students *with basic knowledge of logic*, which was required to understand the proofs. Screening criteria were familiarity with first-order logic (e.g. through a lecture), a stable Internet connection and the permission to record their handwriting and voice during the experiment. One participant was excluded since they did not understand the proofs but rated them as easy. The mean of the participant’s experience with propositional logic was $M = 3.25$ ($SD = 1.0$), on a scale of 1 to 5. Furthermore, 37.5% of the participants seldomly worked with propositional logic, while 31.3% worked with it often.

Hypotheses. We stated three hypotheses concerning the participants’ self-rating of the difficulty of the proofs and their self-rated experience with logic.

Hypothesis 1: It is easier to understand a short, concise explanation than a longer version (in the same representation format).

Hypothesis 2: Users with less experience in logic can understand the longer text better than a short tree proof. This will be shown by a lower difficulty rating of the long textual proof.

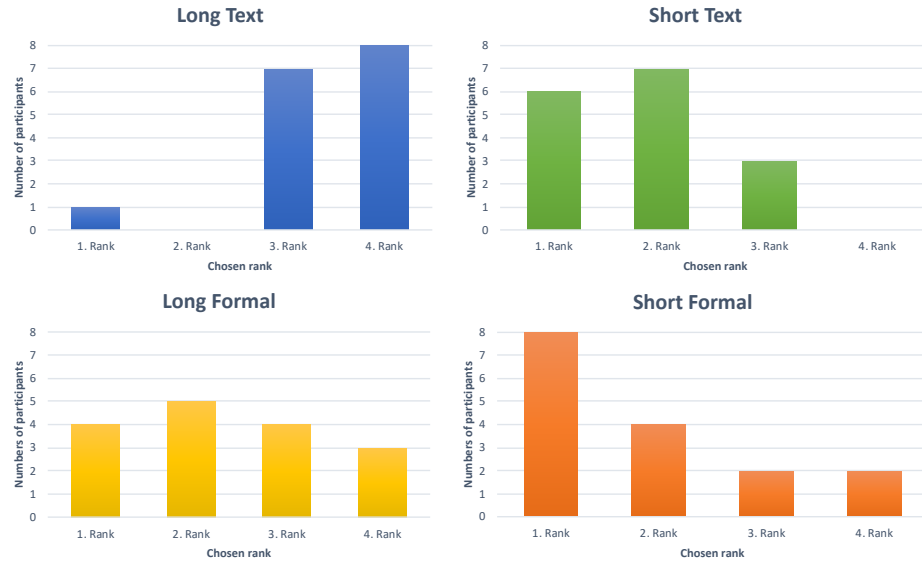


Fig. 2. The participants' ranking of conditions with 1 = very easy and 4 = very difficult

Hypothesis 3: Users with more experience in logic can understand a long tree proof better than a long text. This will be shown by a lower difficulty rating of the long tree proof.

Results. For *Hypothesis 1*, a multiple linear regression with contrast coding (K1, K2, K3) was conducted. K1 contrasted the textual representation against the tree. K2 contrasted the short vs. long proofs and K3 the interaction between the two general conditions. The three contrasts explained 14.2% of variance in the rating after each proof, $R^2 = .14$, $F(3, 60) = 3.30$, $p < .05$. Only K2 was found to be a significant predictor in the linear regression, $\beta = -.29$, $t(60) = -2.42$, $p < .05$. This means that the participants rated the shorter proofs as being easier than the longer ones, which was independent of the presentation format. Thus, *Hypothesis 1* could be supported by our data.

For *Hypotheses 2* and *3*, we computed moderator analyses with the two condition combinations as a predictor, the experience as a moderator variable and the rating after each proof as the criterion. However, neither *Hypothesis 2* nor *3* was supported by our data. Experience with logic did not make a difference on the understanding of the different proof representations.

Additionally to the three hypotheses, we used Friedman's test for comparing the comprehensibility ranking of the proof representations at the end of the experiment (first rank = very easy, fourth rank = very difficult). It revealed a significant difference in the ranking of the condition combinations, $\chi^2(3) = 15.29$, $p < .01$ with a moderate effect size (Kendall's $W = .32$). For the post-hoc pairwise comparisons, Bonferroni correction was used, which resulted in a p -threshold of 0.008, resulting in only two significant comparisons. The participants' ranking of

condition combinations is shown in Figure 2. The combination *Short Text* was preferred over *Long Text*, $Z = 1.53$, $p < .008$. The median ranking for *Short Text* and *Long Text* was 2 and 3.5, respectively. Additionally, *Short Tree* was preferred over *Long Text*, $Z = 1.50$, $p < .008$. *Short Tree* had the lowest median ranking with 1.5. Both comparisons showed moderate effect sizes with $r = 0.38$. The median ranking for *Long Tree* was 2.

Study II – Connecting Cognitive Abilities and Proof Understanding

Our first experiment revealed some weaknesses in our design choices. First, the direct interviews with each person meant that we were only able to include few participants. Therefore, in the following we designed our experiments using automated surveys. Second, the choice of proofs using real domains was not ideal, as sometimes participants immediately spotted axioms that were counter-intuitive, without looking at the proof. This is why we started to use nonsense domains that could not interfere with participants' prior knowledge. Last but not least, the self-rating of experience in logic may be influenced by participants' confidence or a fear of negative evaluation. Thus, we wanted to replace the subjective experience rating by a more objective measure of an individual's ability to understand logical proofs. To evaluate the suitability of standardized tests for our purposes, we conducted the following experiment comparing the International Cognitive Ability Resource (ICAR16⁵) [16] questionnaire against the performance on tasks related to DL proofs.

Design. We used LimeSurvey⁶ for hosting our online survey. Since we did not pre-screen our participants for experience with logic, we included an introduction explaining the structure of proof trees. In order to exclude the effect of tiredness, the order of the ICAR16 questions and the proof tasks was randomized.

Material. To assess the participants' cognitive abilities, the abbreviated form of ICAR16 was applied. It consists of 16 questions equally distributed over four types: matrix reasoning, letter and number series, verbal reasoning, and 3-dimensional rotation. In the end, a mean score was calculated by coding correct answers with 1 and incorrect answers with 0. Thus, the maximum score was 1, while the minimal score was 0. The internal consistency of ICAR16 is $\alpha = .81$ [16].

To test the performance with logical reasoning, participants had to solve two tasks. The first described a set of 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 tree proof that contained a blank node, and they were asked which of some given statements would be valid labels for the node in the context of the proof ("yes", "no", "I do not know"). The score of

⁵ <https://icar-project.com/>

⁶ <https://www.limesurvey.org/>

the performance in both tasks was calculated as the number of correct answers. The highest possible score was 24.

Further Information. As before we asked participants about their experience with propositional logic and their difficulty rating of each task.

Participants (see Table 2). We did not exclude any participants based on the attention checks because no one missed more than one attention check. The mean of the participants' self-reported experience with propositional logic was $M = 1.83$ ($SD = 1.18$), on a scale of 1 to 5. Additionally, 56.4% of the participants had never worked with propositional logic before.

Hypothesis. The only hypothesis was that the ICAR16 score predicts the performance in the logical tasks.

Descriptive Results. The mean of the ICAR16 scores was $M = 0.55$ ($SD = 0.24$) with the participants' performance being spread in a normal distribution. The maximal achieved score was 1, the minimum was 0. The mean of the score for both logical reasoning tasks was $M = 15.99$ ($SD = 3.3$), with the maximum score being 23 and the minimum 6. The performance in these tasks was also normally distributed across the participants.

Regression analysis. 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 [15].

Study III – Logical Abilities and Proof Representation Preferences

We now return to our main research question of which proof representation is more preferred and results in a better performance in certain groups of participants. For this experiment we investigated interactive, static, tree and textual proof formats. Given that ICAR16 scores are highly correlated with performance on logical reasoning tasks, we used it in our next experiment to distinguish participants by their logical ability level. The goal was to find a difference in the (subjective) preferences and (objective) performance on each proof representation, depending on the user's level of logical reasoning ability.

Conditions and Design. We used two different conditions with two levels each. One condition was the proof representation; either tree-shaped or textual. The other condition was the interactivity of the proof representation; either static or interactive. Thus, there were the four following condition combinations: (**ir**) interactive tree, (**sr**) static tree, (**ix**) interactive text, and (**sx**) static text. We again used a 2×2 within-subjects design with a Latin square design. The independent variable in the main study is the ICAR16 score. Objective performance (the number of correct answers) and subjective rating of proofs as well as proof rankings are dependent variables.

The survey was again implemented using LimeSurvey. As in the first experiment, the order of the ICAR16 and the proof question groups was randomized. Moreover, each participant was randomly assigned to one of the four groups according to the Latin square. Before the proof tasks, there was a short explanation and a small training example for both interactive formats (**ir**, **ix**).

Material. We again used ICAR16 to assess the participants’ cognitive abilities.

We developed four artificial proofs of roughly the same difficulty level. The statements of each proof were given in textual form (also for (**ir**, **sr**)) using nonsense words. The (**ir**) version started with only the final conclusion visible, and participants could interact with each node to reveal or hide its predecessors in the tree. The (**ix**) worked in a different way. At the beginning, participants saw only the first sentence, i.e. the first assumption. They could then reveal the next sentences step-by-step, and also highlight the premises that were used to obtain a selected statement. Moreover, both interactive representations, (**ir,ix**), could be freely zoomed and panned. The interactive proofs were provided by a prototypical web application ⁷ for explaining DL entailments [2, 6, 19]. For this study, it was extended by a (linear) textual representation of proofs. The modes of interaction were kept relatively basic to avoid overwhelming participants who had little experience with logic and proofs.

For each proof, there were three questions. Each question had 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. Thus, the highest possible score was 12.

Further Information. We again asked participants about the experience with propositional logic and the difficulty rating of proofs, as well as a ranking of all four conditions they had seen according to their relative comprehensibility.

Participants (see Table 2). The mean of the participants’ experience with propositional logic was $M = 1.76$ ($SD = 1$) on a scale between 1 and 5. Furthermore, 60.7% of the participants indicated that they had never worked with propositional logic. Due to technical errors, the proofs were not displayed for 3 participants, which were excluded. Four attention checks were implemented in the study. 13 participants with more than 2 incorrectly answered attention checks were excluded from the analysis.

Hypotheses. We stated two hypotheses concerning the preferences and performance differences between the proof representations.

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.

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

⁷ <https://imld.de/evonne>

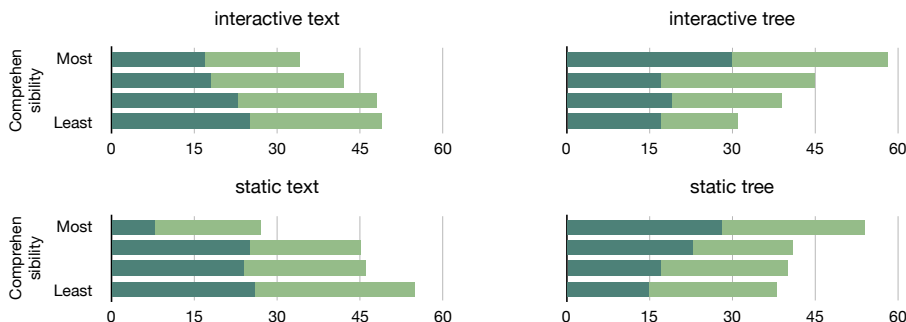


Fig. 3. Rankings of all 173 participants (light bars) and of the 83 participants with high ICAR scores (dark bars) for each condition combination.

in performance and difficulty rating between the condition combinations and in the final comprehensibility ranking, in dependence of the ICAR16 scores.

Results. After the assumptions were considered as tenable, a regression analysis was carried out, to confirm the results of Study II. Again, the predictive effect of the ICAR16 on the performance in the proofs was significant, $F(1, 171) = 24.8$, $p < .001$. With an $R^2 = .13$ (corrected $R^2 = .12$), the model shows a moderate explained variance (Cohen, 1988).

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.

For ICAR16 the mean was $M = 0.46$, while it was $M = 2.36$ for the proof performance. The group containing those participants who scored low in the ICAR16 achieved $M = 1.9$ across all proofs. In contrast, the group of participants with high ICAR16 scores showed an overall proof performance of $M = 2.87$.

Performance and Comprehensibility Ratings. To compare the proof performance and the subjective comprehensibility ratings after each proof, we ran a multivariate analysis of variance (MANOVA). All the assumptions were considered as tenable. We found no significant overall difference between the conditions across the two ICAR groups, Pillai's Trace = .01, $F(6, 1376) = 1.41$, $p = .206$. Also when looking at the groups separately, we could not find any significant differences between the representations, neither in the low-ICAR group (Pillai's Trace = .03, $F(6, 712) = 1.90$, $p = .078$) nor in the group with high scores (Pillai's Trace = .01, $F(6, 656) = .53$, $p = .788$). Thus, we could not detect differences in the comprehensibility ratings as well as the performance between the various representations in each cognitive ability group and across the two groups.

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 ICAR groups, $\chi^2(3) = 17.16$, $p = .001$, $n = 173$ (see Figure 3, light bars). Post-hoc pairwise comparisons were Bonferroni-corrected

and showed three significant comparisons. The **(ir)** was significantly more often ranked higher than the **(ix)** ($z = .40$, $p = .024$, Cohen’s effect size $r = .03$) and also higher than static text ($z = -.50$, $p = .002$, Cohen’s effect size $r = .04$). The **(sr)** representation was also ranked significantly higher than **(sx)**, $z = .39$, $p = .032$, Cohen’s effect size $r = .03$ (see Figure 3).

A Friedman’s test in the group with high ICAR performance showed a significant difference in the ranking of representations, $\chi^2(3) = 12.73$, $p = .005$, $n = 83$ (see Figure 3, dark bars). Bonferroni-corrected post-hoc pairwise comparisons revealed two significant comparisons. There is a significant difference between **(sr)** and **(sx)** ($z = .59$, $p = .019$, Cohen’s effect size $r = .06$) with **(sr)** being ranked higher than **(sx)**. The **(ir)** was also preferred before **(sx)**, ($z = -.54$, $p = .041$, Cohen’s effect size $r = .06$).

The low-ICAR-performers showed no significant difference in the ranking of representations, $\chi^2(3) = 6.70$, $p = .082$, $n = 90$.

Study IV – Final Experiment

The main shortcoming of the previous experiment was the difficulty of the proof tasks, which could be seen in the mean score of 2.36 out of 12. Therefore, we designed another experiment where the difficulty of the proof tasks was adjusted. We also did not include the interactive conditions to be able to focus more on the difference between the text vs. tree proofs. Furthermore, the number of proof tasks was reduced and the nonsense words were replaced by letters, to reduce the cognitive overload that some participants had reported in the previous study.

Conditions and Design. We only considered one condition with two levels, namely textual and tree-shaped proof representation. We also used a between-subjects design, which means that each participant saw either only text proofs or only tree proofs. Dependent variables were proof performance and subjective comprehensibility rating. The independent variable was the ICAR16 score.

We conducted the experiment via LimeSurvey. As before, the order of ICAR16 and the proof tasks was randomized. Each participant was randomly assigned to one of the two groups. We again included a short training example at the beginning of the proof tasks.

Material. We again used ICAR16 to assess the participants’ cognitive abilities (see page 7). For the proof tasks, we used simplified versions of the proofs from the previous study, where additionally the nonsense words were replaced by letters, e.g. “every G has at least two Ys”. Overall, there were 2 proofs with 3 questions each. Thus, the highest possible score was 6.

Further Information. We again collected information about participants’ experience with propositional logic and subjective ratings after each proof.

Participants (see Table 2). We excluded 7 participants because they did not pass the two attention checks. The experience with propositional logic was $M = 1.53$ ($SD = 0.97$). Moreover, 69.4% indicated that they had never worked with propositional logic before.

Hypotheses. We again wanted to test our previous hypothesis that the comprehensibility of a tree-shaped vs. textual proof depends on the cognitive abilities. This would be shown by a difference in performance and difficulty rating between the conditions, in dependence of the ICAR16 scores.

Results. The mean of ICAR16 was $M = 0.36$ ($SD = 0.20$) while it was $M = 2.30$ ($SD = 1.25$) for the proof performance. We carried out a regression analysis to confirm the results of the previous two studies. These results should indicate that ICAR16 scores predict proof performance. This is a precondition for any following analyses, because we cannot split the sample based on the ICAR16 values if they are not sufficiently related to the proof values. The predictive effect of the ICAR16 on the proof performance was not significant, $F(1, 106) = 2.26$, $p = .135$, which is why we did not perform any further tests.

General Discussion

Our main hypotheses that experience with logic or logical ability influences the subjective rating or objective performance on different proof representations could not be confirmed (see Hypotheses 2 and 3 in Study I, Hypothesis 2 in Study III and the only hypothesis in Study IV). This may be partially due to the shortcomings of each of the experiments, which we discuss in more detail below. In addition, we could not find any advantage of specific representations when it comes to the performance on proof-related tasks, even when ignoring the ICAR16 scores (see Studies III and IV).

Nevertheless, our first experiment clearly showed a preference for shorter proofs based on the subjective difficulty ratings and relative rankings of the conditions by the participants. This shows that it is worthwhile to investigate techniques for automatically shortening proofs to remove easy or redundant steps that only distract the users. As a side result, in the second experiment we demonstrated that cognitive abilities as measured by the standardized ICAR16 questionnaire can be used as a predictor for the performance on logical reasoning tasks. The final ranking in third experiment showed a further subjective preference for the conditions with tree-shaped proofs over their textual counterparts, but this did not seem to impact the objective performance measure nor the subjective ratings the participants gave after each proof. These preferences were largely driven by the group with higher ICAR16 performance (cf. Figure 3).

Limitations

A general shortcoming of our main hypothesis was perhaps that it was too specific. If there are any effects of proof representation between user groups, they were maybe too small to detect in our experiments. After the first study, we recruited more participants through the online platform Prolific, but this also came with a loss of quality in the responses that we could not completely control with the attention checks. Since everyone was paid the same amount of money, the goal of many participants was to complete the study as fast as possible. Several

participants even finished the larger studies (including both ICAR16 and proof tasks) with successful attention checks in under 15 minutes, which hints at a loss of quality in the responses. A solution to this could be using open instead of multiple-choice questions. However, such answers must be evaluated manually by an expert according to a-priori fixed criteria.

Another limitation of the first study was also that it did not include an objective measure of performance; participants were simply asked to describe their process of understanding the proofs which was later rated by an expert. We therefore included objective proof tasks in Study III, which however were too hard for most of the participants. According to the aims of our study, we did not pre-select participants according to their experience with logic or field of studies. 55.5% of the participants had no experience with propositional logic and 60.7% had never worked with it. For many participants, even the ones with higher ICAR scores, the proof tasks were very challenging, resulting in a mean score of $M = 2.36$ out of a total of 12. 15 people commented about the high difficulty level in the end, and only 3 said the proofs were easy to understand. This resulted in many data points being clustered on the lower end of the scale and differences being more difficult to detect.

In general, a between-subjects design is better suited to show differences between proof representations because there is no interference between the conditions, but this requires even more participants. In the last study, we attempted to do this and also adjusted the difficulty of the proofs. Unfortunately, this study failed to exhibit even the strong connection between ICAR16 scores and proof performance that had been shown by the previous two studies. Possible reasons for this are that there were too few data points for the proof tasks (the maximal score was 6 since we did not want to overload the participants) and that the participants in general seemed to differ from previous studies. It seemed that participants showed higher ICAR16 scores in the second ($M = 0.55$, $SD = 0.24$) and third ($M = 0.46$, $SD = 0.24$) than in the fourth study ($M = 0.36$, $SD = 0.20$), and the self-reported experience with logic followed a similar pattern. This could be a reason why the ICAR16 scores did not predict the proof performance in Study IV.

Future work

Although several of the experiments indicate a subjective preference of tree proofs over texts, we would like to study more formally whether this can also influence performance (independent of the membership to any particular user group such as logic experts or people with high cognitive abilities). In that context, it could also make a difference whether the individual statements in tree proofs are shown as natural language sentences or using logical syntax (as in our first study). Another question with a larger expected effect is whether showing proofs actually makes a difference when compared to only showing justifications, i.e. the premises/leaves of the tree proofs without intermediate inference steps.

Moreover, it would be promising to look at an ontology that is actively used in practice and to study domain experts performing specific relevant explanation

tasks for this ontology. Ultimately, our studies are just a first step towards developing a user-centered interactive explanation tool for DL ontologies. Such a tool should also take into account individual differences, such as user preferences or the user's existing knowledge, e.g. in the form of a *background ontology* that the user is assumed to understand intuitively without needing an explanation.

Acknowledgements This work was supported by the DFG grant 389792660 as part of TRR 248 – CPEC (<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: ISWC (2017). https://doi.org/10.1007/978-3-319-68288-4_2
2. Alrabbaa, C., Baader, F., Borgwardt, S., Dachselt, R., Koopmann, P., Méndez, J.: Evonne: Interactive proof visualization for description logics (system description). In: IJCAR (2022). https://doi.org/10.1007/978-3-031-10769-6_16
3. Alrabbaa, C., Baader, F., Borgwardt, S., Koopmann, P., Kovtunova, A.: Finding small proofs for description logic entailments: Theory and practice. In: LPAR-23 (2020). <https://doi.org/10.29007/nhpp>
4. Alrabbaa, C., Baader, F., Borgwardt, S., Koopmann, P., Kovtunova, A.: On the complexity of finding good proofs for description logic entailments. In: DL Workshop (2020), <http://ceur-ws.org/Vol-2663/paper-1.pdf>
5. Alrabbaa, C., Baader, F., Borgwardt, S., Koopmann, P., Kovtunova, A.: Finding good proofs for description logic entailments using recursive quality measures. In: CADE (2021). https://doi.org/10.1007/978-3-030-79876-5_17
6. Alrabbaa, C., Baader, F., Dachselt, R., Flemisch, T., Koopmann, P.: Visualising proofs and the modular structure of ontologies to support ontology repair. In: DL Workshop (2020), <http://ceur-ws.org/Vol-2663/paper-2.pdf>
7. Alrabbaa, C., Borgwardt, S., Knieriem, N., Kovtunova, A., Rothermel, A.M., Wiehr, F.: In the hand of the beholder: Comparing interactive proof visualizations. In: DL Workshop (2021), <http://ceur-ws.org/Vol-2954/paper-2.pdf>
8. Androutsopoulos, I., Lampouras, G., Galanis, D.: Generating natural language descriptions from OWL ontologies: The NaturalOWL system. JAIR **48** (2013). <https://doi.org/10.1613/jair.4017>
9. Baader, F., Brandt, S., Lutz, C.: Pushing the \mathcal{EL} envelope. In: IJCAI (2005), <http://ijcai.org/Proceedings/09/Papers/053.pdf>
10. Baader, F., Horrocks, I., Lutz, C., Sattler, U.: An Introduction to Description Logic. Cambridge University Press (2017). <https://doi.org/10.1017/9781139025355>
11. Baader, F., Peñaloza, R., Suntisrivaraporn, B.: Pinpointing in the description logic \mathcal{EL}^+ . In: KI (2007). https://doi.org/10.1007/978-3-540-74565-5_7
12. Baader, F., Suntisrivaraporn, B.: Debugging SNOMED CT using axiom pinpointing in the description logic \mathcal{EL}^+ . In: KR-MED (2008), <http://ceur-ws.org/Vol-410/Paper01.pdf>
13. Borgida, A., Franconi, E., Horrocks, I.: Explaining \mathcal{ALC} subsumption. In: ECAI (2000), <http://www.frontiersinai.com/ecai/ecai2000/pdf/p0209.pdf>

14. Borgwardt, S., Hirsch, A., Kovtunova, A., Wiehr, F.: In the Eye of the Beholder: Which Proofs are Best? In: DL Workshop (2020), <http://ceur-ws.org/Vol-2663/paper-6.pdf>
15. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates (1988). <https://doi.org/10.4324/9780203771587>
16. Condon, D.M., Revelle, W.: The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence* **43** (2014). <https://doi.org/10.1016/j.intell.2014.01.004>
17. Donadello, I., Dragoni, M., Eccher, C.: Explaining reasoning algorithms with persuasiveness: a case study for a behavioural change system. In: ACM Symposium on Applied Computing (2020). <https://doi.org/10.1145/3341105.3373910>
18. Engström, F., Nizamani, A.R., Strannegård, C.: Generating comprehensible explanations in description logic. In: DL Workshop (2014), http://ceur-ws.org/Vol-1193/paper_17.pdf
19. Flemisch, T., Langner, R., Alrabbaa, C., Dachsel, R.: Towards designing a tool for understanding proofs in ontologies through combined node-link diagrams. In: VOILA Workshop (2020), <http://ceur-ws.org/Vol-2778/paper3.pdf>
20. Hayes, A.F.: Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. Guilford Publications (2017). <https://doi.org/10.1111/jedm.12050>
21. Horridge, M.: Justification Based Explanation in Ontologies. Ph.D. thesis, University of Manchester, UK (2011), https://www.research.manchester.ac.uk/portal/files/54511395/FULL_TEXT.PDF
22. Horridge, M., Bail, S., Parsia, B., Sattler, U.: Toward cognitive support for OWL justifications. *Knowl.-Based Syst.* **53** (2013). <https://doi.org/10.1016/j.knosys.2013.08.021>
23. Horridge, M., Parsia, B., Sattler, U.: Justification oriented proofs in OWL. In: ISWC (2010). https://doi.org/10.1007/978-3-642-17746-0_23
24. IBM: SPSS Statistics, <https://www.ibm.com/products/spss-statistics>
25. Kalyanpur, A.: Debugging and Repair of OWL Ontologies. Ph.D. thesis, University of Maryland, College Park, USA (2006), <http://hdl.handle.net/1903/3820>
26. Kazakov, Y., Klinov, P., Stupnikov, A.: Towards reusable explanation services in Protege. In: DL Workshop (2017), <http://www.ceur-ws.org/Vol-1879/paper31.pdf>
27. Kazakov, Y., Krötzsch, M., Simancik, F.: The incredible ELK – from polynomial procedures to efficient reasoning with \mathcal{EL} ontologies. *JAR* **53** (2014). <https://doi.org/10.1007/s10817-013-9296-3>
28. Kontopoulos, E., Bassiliades, N., Antoniou, G.: Visualizing semantic web proofs of defeasible logic in the DR-DEVICE system. *Knowl. Based Syst.* (2011). <https://doi.org/10.1016/j.knosys.2010.12.001>
29. Kuhn, T.: The understandability of OWL statements in controlled English. *Semantic Web* **4** (2013). <https://doi.org/10.3233/SW-2012-0063>
30. McGuinness, D.L.: Explaining Reasoning in Description Logics. Ph.D. thesis, Rutgers University, NJ, USA (1996). <https://doi.org/10.7282/t3-q0c6-5305>
31. Meehan, T.F., Masci, A.M., Abdulla, A., Cowell, L.G., Blake, J.A., Mungall, C.J., Diehl, A.D.: Logical development of the cell ontology. *BMC Bioinformatics* **12** (2011). <https://doi.org/10.1186/1471-2105-12-6>
32. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. *AI* **267** (2019). <https://doi.org/10.1016/j.artint.2018.07.007>

33. Nguyen, T.A.T., Power, R., Piwek, P., Williams, S.: Predicting the understandability of OWL inferences. In: ESWC (2013). https://doi.org/10.1007/978-3-642-38288-8_8
34. Schiller, M.R.G., Glimm, B.: Towards explicative inference for OWL. In: DL Workshop (2013), http://ceur-ws.org/Vol-1014/paper_36.pdf
35. Schiller, M.R.G., Schiller, F., Glimm, B.: Testing the adequacy of automated explanations of EL subsumptions. In: DL Workshop (2017), <http://ceur-ws.org/Vol-1879/paper43.pdf>
36. Schlobach, S.: Explaining subsumption by optimal interpolation. In: JELIA (2004). https://doi.org/10.1007/978-3-540-30227-8_35
37. Schlobach, S., Cornet, R.: Non-standard reasoning services for the debugging of description logic terminologies. In: IJCAI (2003), <http://ijcai.org/Proceedings/03/Papers/053.pdf>
38. Schulz, S.: The role of foundational ontologies for preventing bad ontology design. In: BOG Workshop (2018), http://ceur-ws.org/Vol-2205/paper22_bog1.pdf
39. Simancik, F., Kazakov, Y., Horrocks, I.: Consequence-based reasoning beyond Horn ontologies. In: IJCAI (2011). <https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-187>