In the Heart of the Beholder: User-Tailored Explanations for Description Logics

Stefan Borgwardt[1], Anke Hirsch[2], Nina Knieriemen[2] and Alisa Kovtunova[1]

[1]Technische Universität Dresden, Faculty of Computer Science, Dresden, Germany
[2]Saarland University, Saarland Informatics Campus and DFKI, Saarbrücken, Germany

Abstract

Many techniques have been developed to explain logical reasoning, such as proofs or abduction. However, such methods are useful mainly for experts in logic, e.g., for debugging ontologies. For actually explaining logical consequences and missing consequences to end users of logic-based systems, e.g., in the Semantic Web, it is necessary to study how to adapt and present such explanations in an understandable way. We report on a series of user studies investigating both entailment explanations via proofs and missing entailment explanations via abduction and counterexamples in the context of Description Logic (DL) ontologies. For the former, we assess the influence of prior knowledge on the necessary explanation granularity; for the latter, we compare the efficacy of abductive versus counterexample-based approaches. While we did not find objectively quantifiable results, we analyse and discuss the results of detailed qualitative user interviews, and extract recommendations for executing user studies on logical reasoning systems.

Keywords

user studies, formal proofs, visualisation

1. Introduction

Symbolic techniques in Artificial Intelligence are often portrayed as inherently more explainable, although less flexible, than approaches based on machine learning. Nevertheless, actually *explaining* the output of, say, a DL reasoner, still requires the development of dedicated techniques. For DL experts, many methods have been proposed to explain *entailments*—via justifications [1, 2, 3], interpolation [4], provenance [5, 6] and proofs [7, 8, 9, 10, 11, 12]—as well as *non-entailments*—by abduction [13, 14, 15, 16, 17] and counterexamples [18, 19, 20]. However, full explanations may be long and technical, and thus not always suited for end users of AI systems with symbolic components. Therefore, efforts have been made to study the cognitive complexity and interpretability of various forms of user-adapted explanations [21, 22, 23, 24, 25, 26, 27, 28, 29, 30].

Such studies can be generally classified into *quantitative* and *qualitative* studies. While the former kind aims to answer research questions such as "Do the explanations help the users in their work?" by measuring quantifiable outcomes on concrete tasks, e.g., the time taken to repair an ontology, qualitative studies ask more subjective questions like "What are the relative advantages and disadvantages of different kinds of explanations?", and often employ open-ended questions and user interviews. While the goal of qualitative studies is to verify the research questions by means of statistical tests, the outcome of quantitative studies is a distillation of discussions with the study subjects, which can give more insight into the users' thought processes, but does not provide objective answers [31, 32]. In previous work with co-authors [28] on explaining entailments by using proofs, we have already reported on the difficulties of drawing quantitative conclusions from user studies, regardless of scale (number of participants), mode (online interviews or questionnaires), user group (DL researchers or random internet users), logical syntax (DL syntax or text), and domain (real concept/role names or made-up

DL 2026: 39th International Workshop on Description Logics, July 17–19, 2026, Lisbon, Portugal
stefan.borgwardt@tu-dresden.de (S. Borgwardt); anke.hirsch@uni-saarland.de (A. Hirsch); nina.knieriemen@dfki.de (N. Knieriemen); alisa.kovtunova@tu-dresden.de (A. Kovtunova)
https://lat.inf.tu-dresden.de/~stefborg/ (S. Borgwardt); https://lat.inf.tu-dresden.de/~alisa/ (A. Kovtunova)
0000-0003-0924-8478 (S. Borgwardt); 0009-0008-8377-8905 (A. Hirsch); 0000-0001-9936-0943 (A. Kovtunova)

words or just letters). Nevertheless, our investigations clearly showed a preference for shorter proofs and tree-shaped over text-based proofs for explaining entailments. These preferences were stronger in participants with higher aptitude for logic-based reasoning tasks.

In this paper, we report on a series of subsequent user studies with different designs concerning both entailment explanations and missing entailment explanations, where we focus on two research questions. First, we wanted to compare the efficacy of explanations for *non-entailments* based on abduction and counterexamples (see Section 2). These two approaches address missing entailments in complementary but distinct ways: the latter directly shows why an entailment fails, while the former suggests how it could be made to succeed. Though not equivalent, both can provide useful pointers for clarifying why an entailment does not follow. What remained an open question, however, was whether either approach confers a distinct advantage for the user. Our goal was therefore to determine which formalism is better suited for developing intuitive tools to explain missing consequences. Additionally, in Section 4 we investigated how prior experience with the domain of an ontology affects the level of detail a (proof-based) explanation needs to provide. Our underlying motivation was to determine whether explanations can be improved by hiding subproofs that are already known to the user, thereby making the overall experience more effective and less cognitively demanding. These two quantitative experiments did not yield statistically significant results or clear descriptive patterns. However, we followed them up with qualitative studies in Sections 3 and 5 to gain more insights into the thought processes of the users when dealing with these kinds of explanations, in particular when using a real ontology engineering environment such as Protégé [29].[1] These qualitative studies revealed two descriptive patterns, reported as exploratory observations rather than statistically tested hypotheses. For missing entailments, most participants valued explanation support, with a slight preference for abduction over counterexamples. With respect to the second research question, the responses on perceived difficulty suggest that prior domain experience indeed facilitates proof understanding, with experienced users reporting lower difficulty than novices. We provide a detailed analysis and discussion of our findings in the corresponding sections.

Referring to our previous studies as Studies I–IV [28], and the ones presented in this paper as Study V–VIII, we conclude with a summary in Table 2 with recommendations for executing user studies in the area of knowledge representation and reasoning.

More details and printable versions of the surveys are available online [33].

2. Study V – Comparing the Comprehensibility of Explanations for Missing Logical Entailments

While reasoning can sometimes reveal unexpected entailments that require explanation, the problem often lies with what is *not* entailed. To explain missing entailments and suggest how to repair them, two primary approaches exist: counterexamples and abduction. A counterexample is a model of the ontology that does not satisfy the target entailment. In this work, we focus specifically on the part of the model relevant to the missing entailment [19, 20]. In abduction, the missing entailment is explained by sets of axioms that, when added to the ontology, enable it to entail the missing consequence. To generate such sets in this section and Section 3, we used the tools CAPI[2] [14] and LETHE-abduction[3] [13].

We wanted to investigate which one (if any) of the methods, counterexample or abduction, can facilitate the comprehension of why an entailment is missing. The study design and hypotheses were preregistered on the Open Science Framework [34].

Procedure The study was conducted online via a questionnaire in LimeSurvey in German. After reading and consenting to data privacy terms and regulations, participants answered demographic questions and two questions about their experience with propositional and first-order logic. Participants

[1] https://protege.stanford.edu/
[2] https://lat.inf.tu-dresden.de/~koopmann/CAPI/
[3] https://github.com/PKoopmann/LETHE-0.8

were randomly assigned to one of three groups, seeing either no explanations (N), counterexamples (C) or sets of axioms generated by abduction (A). A training phase (see Figure 7 in Appendix A) was followed by instructions of the task. Each task consisted of a small ontology and a missing entailment, followed by an explanation for Groups (C) and (A). In order to reduce the influence of participants' background knowledge, we used two commonly known domains (pizzas and movies). Before the main task, an example was presented (see Figures 8, 9 in Appendix A). Participants rated how well they understood that the statement does not follow (from 1 = not at all to 5 = very well). Participants in Groups (C) and (A) also indicated how helpful the explanation was (from 1 = not at all helpful to 5 = very helpful). In the main phase (see Figure 10 in Appendix A), each participant completed three tasks involving common types of missing entailments: a missing universal quantifier ($\exists r.A \not\sqsubseteq \forall r.A$),[4] non-interaction of two existential quantifiers ($\exists r.A \sqcap \exists r.B \not\sqsubseteq \exists r.(A \sqcap B)$), and a missing implication inside an existential quantifier ($\exists r.A \not\sqsubseteq \exists r.B$ due to $A \not\sqsubseteq B$). After seeing an example missing entailment (with an explanation in Groups (C) and (A)), they had to determine whether a list of similar statements follows from the ontology or not. Performance was measured by the number of correct answers. Participants then indicated how helpful the example was to solve the statements from 1 = not at all helpful to 5 = very helpful.

Hypotheses:

1. There is a difference in performance between abduction and no explanation.
2. There is a difference in performance between counterexamples and no explanation.
3. There is a difference in performance between abduction and counterexamples.

Sample We used the software program G*Power [35, 36, 37] to conduct an a priori sample size calculation for a statistical ANOVA test (analysis of variance). With a power of .80 to detect a medium effect size of f = .25 at the standard α = .05 error probability, with 3 groups (df = 2), a total sample size of 158 was calculated. Since participants were expected to have a basic understanding of propositional logic, they were recruited from the "Formale Systeme" lecture at TU Dresden, which is part of the computer science Bachelor's and Diploma programmes and attended by 447 students. In the end, 63 participants took part in our study. They were between 18 and 37 years old, with a mean age of 21.79 years (*SD* = 3.93). Ten participants were female, 52 were male and one person preferred not to disclose. With the participation, they had the chance to win one out of 10 Amazon vouchers worth 20 €. Winners were selected at random after the survey. Participation took about 40 minutes. Participants were asked to assess the quality of their own data. As no participant indicated low data quality, none were excluded on this basis. On a scale from 1 = none to 5 = expert, participants indicated their experience with propositional logic with a mean of 3.41 (SD = 0.71). For first-order logic, mean experience was 2.78 (SD = 0.89).

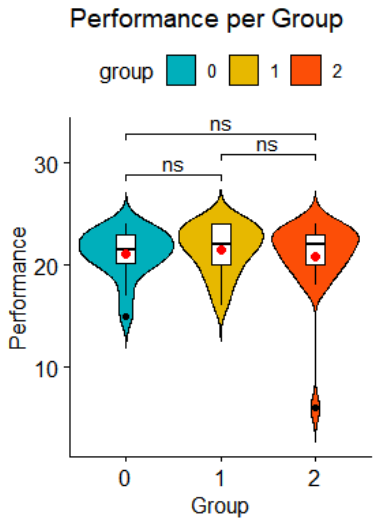
Results Analyses were calculated using R [38]. To evaluate potential differences in performance across the three conditions, we calculated a performance score by summing the number of correctly answered tasks for each participant with a maximum possible score of 24. Of the participants, 18 were assigned to Group (N), 24 to Group (C) and 21 to Group (A). An ANOVA was conducted to test our hypotheses with performance as the dependent variable and group assignment as the independent variable. The Shapiro–Wilk test was significant, indicating a violation of the normality assumption. Since the ANOVA is relatively robust to violations of this assumption [39, 40], we nevertheless performed a parametric test. In a Welch ANOVA, no significant differences were found comparing the three groups regarding performance ($F(2, 37) = 0.32$, $p = .731$, $\eta_p^2 = .003$). The Levene's test was not significant ($F(2, 60) = 0.47$, $p = .629$), indicating that variance homogeneity is not violated. Thus, to evaluate our specific hypotheses, post hoc comparisons were calculated using Tukey's test. Regarding the first hypotheses,

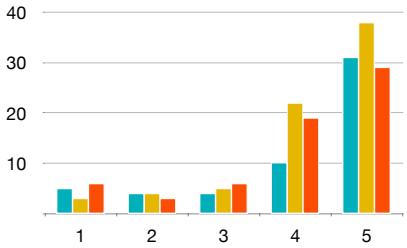
[4] For example, the conclusion that "Pizza Margherita is vegetarian" does not follow solely from the ingredients it contains (expressed with $\exists$) and the general rule that vegetarian products contain only vegetarian ingredients (an axiom with $\forall$). What is *missing* is the additional statement that Margherita has no other ingredients.

Table 1 & Figure 1

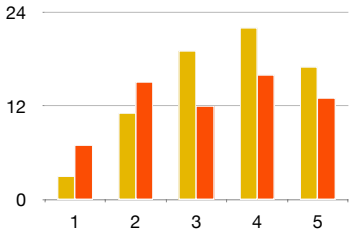
Descriptive statistics (left) and a violin plot (right) showing group differences in performance. The figure includes box plot (interquartile range), end of line = min / max, rotated kernel density plot and means displayed as red dots.

Group	Mean (SD)	Min	Max
0: No explanation	21.17 (2.36)	15	24
1: Counterexample	21.54 (2.38)	16	24
2: Abduction	20.81 (3.87)	6	24

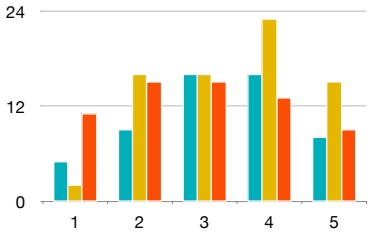




(a) How well did you understand that the statement does not follow?



(b) How helpful was the explanation?



(c) How helpful was the example before to solve the main task?

Figure 2: Answers per group to the subjective questions questions after a task, where 1 = not at all, 5 = very well. Colors indicate groups in the same way as in Figure 1: cyan = no explanation, orange = counterexample, red = abduction.

there was no significant difference in performance between abduction and no explanation ($t(60) = 0.38$, $p = .925$, $d = 0.12$). Further, there was no significant performance difference between Groups (C) and (N) ($t(60) = -0.41$, $p = .913$, $d = -0.13$). Regarding our last hypothesis, there was also no significant difference in performance between abduction and counterexamples ($t(60) = 0.73$, $p = .686$, $d = 0.25$). Figure 1 shows a violin plot displaying group differences. Descriptive values of the groups can be found in Table 1.

Explorative Analysis Regarding the subjective ratings of understanding and helpfulness, we looked at mean differences to explore the data further. Descriptive statistics can be seen in Figure 2. The figure summarizes ratings for the three subjective questions. Overall, ratings were relatively similar across groups, consistent with the lack of statistically significant differences in performance. However, small descriptive differences could indicate minor variations in user preferences. In particular, counterexamples received slightly higher ratings across all measures, including understanding (see Figure 2a), helpfulness of the explanation (see Figure 2b), and helpfulness of the example (see Figure 2c). The other conditions showed similar but marginally lower ratings.

Discussion In this study, a ceiling effect was observed: In all groups, the mean performance score was high and the maximum score was reached by some participants. This suggests that the tasks were likely too easy for the target population, limiting the ability to detect meaningful differences between groups or conditions. Consequently, the measures used may not have been sensitive enough to capture variations in performance. Another limitation is the sample size, which was smaller than originally planned. The inability to recruit the intended number of participants reduces the statistical power of the study and limits the generalizability of the findings. Future research should aim to include more

challenging tasks and ensure adequate sample sizes to provide more robust results.

3. Study VI – Explaining Missing Entailments for OWL Ontologies with Evee

In this section, we summarize a follow-up user study published at KR'24 [29]. Changing the study design, we conducted a qualitative study with regular users of Protégé [41] in the form of online interviews using the think-aloud protocol, intended to get feedback on the usefulness of abduction and counterexample explanations in a realistic ontology engineering setting. The explanations were implemented as Protégé plugins in the Evee project.[5] The study details are available on Zenodo [42].

Procedure During the interviews, participants underwent extensive training that covered relevant features of Protégé and the counterexample and abduction plug-ins, as well as solving training tasks. Participants then performed three tasks and answered questions related to each task. Each task specified a missing entailment in a commonly known domain (pizzas, animals, university lectures), and participants were asked to fix the ontology to achieve the entailment, like $\mathsf{AmericanHot} \sqsubseteq \mathsf{SpicyPizza}$ (see Figure 11 in Appendix A), by extending or adapting the existing axioms about AmericanHot and related classes. Each participant worked on the given ontology in three phases, each with different missing entailments: using no explanation, using only a counterexample, and using only abduction results. After each section, participants were asked which aspects they found especially easy or hard to do, and whether they thought that the explanation service was useful for fixing the ontologies. At the end, participants provided feedback by answering the following questions:

- Did you think the explanation services were useful at all, especially compared to having no explanation service?
- Which of the services do you think was most/least useful? Why?
- Which particular features of the services did you like or not like? Why?

We ran a pilot study with four participants before conducting interviews with 18 participants (3 female, 15 male) between 23 and 74 years old, with a mean age of 37 (SD = 13.53), for the main study.

Results For the data analysis, we transcribed the recorded audio and then used inductive coding [43]: Based on the collected answers, three experts (coders) agreed on short codes, such as "useful for learning OWL/DL", and then independently assigned each participant to the appropriate codes based on their answers. The inter-coder reliability was calculated from the codes of the three experts: The agreement between the coders was sufficiently high for the results to be interpretable. Overall, most of the 18 participants preferred using at least one of the explanation services for the ontology repair tasks, compared to having no explanation support (16 participants for abduction and 15 for counterexamples). Ten participants preferred the abduction service over the counterexamples, whereas six said the opposite. We were also able to collect many useful suggestions for improving the plug-ins, some of which have since been implemented in Evee.

4. Study VII – "I Already Know That": Can Explanations for Experts Be Shortened?

Our next study focused on proofs as explanations for entailments, in particular, the personalisation of logical proofs with respect to a user's prior knowledge. Existing approaches provide static explanations that do not distinguish between different types of users. As a result, novices may feel overwhelmed by large proofs, while experts may find the details redundant. Therefore, we investigated the potential

[5] https://github.com/de-tu-dresden-inf-lat/evee

for enhancing the efficacy of proof explanations through personalisation—specifically, by adapting the level of detail to an individual's level of domain knowledge.

To investigate this, we conducted a within-subjects experiment examining how prior knowledge affects the process of understanding logical proofs. We employed Evonne[6] [44], a web-based interactive tool for visualizing logical reasoning, as our experimental platform. Crucially for the manipulation of dividing participants into experts and non-experts, each participant engaged with two distinct proofs in a counterbalanced design: for one proof, they are positioned as a domain "expert" by receiving targeted background instructions via a video; for the other, they approach it as a "novice" without this preparatory knowledge. This design allows us to compare each individual's performance and experience across the two knowledge conditions, controlling for inherent aptitude differences. The study design and hypotheses were preregistered [45].[7]

Hypotheses Experts need no or less explanation concerning their field of expertise [46]. Thus, our core hypothesis is that when interacting with a proof within their primed domain of expertise, participants will require less explanatory support and find the proof more accessible. We test this through three specific predictions:

- Experts will exhibit more efficient exploration, indicated by a lower click count within the Evonne interface.
- Experts will perceive a proof as easier to understand (rated on a Likert-like rating scale).
- Experts will need less time for proof exploration

Sample We used a sensitivity analysis in G*Power to compute the effect size for a fixed number of participants, because we had limited resources for the study [35, 36, 37]. With 50 participants in the expert and 50 in the novice groups, a test power of .80 and at the standard .05 alpha error probability, we aimed for 100 participants. The analysis suggested that, with 100 participants, we would be able to detect an effect of a medium effect size of f = .25. An effect size of f =.25 corresponds to a medium effect, suggesting a practically meaningful but not large difference between conditions.

We recruited 102 participants via Prolific,[8] all at least 18 years old and fluent in English. They were paid 7£ for participation. During the tasks, participants studied proofs and completed a 5-question cloze test (fill-in-the-blanks with 3 choices each) to verify comprehension of the core argument. Participants who failed 3 or more questions on any task were excluded due to poor data quality. We also excluded participants with incomplete data. After excluding, we retained 95 participants in the dataset. The demographic analysis showed that participants were between 20 and 60 years old, M = 30.2 (SD = 7.95). The sample included 48 women, 45 men, and two non-binary participants.

Procedure The study was conducted online via a questionnaire in LimeSurvey and Evonne. In particular, the code of Evonne and its interface were modified to visualise natural language sentences, host the study proofs online, and log the number of clicks per user.

After participants were informed about the general goal of the study and given their informed consent to the data protection, they provided basic demographic information about age and gender. Before the main tasks, participants completed a brief training phase to become familiar with the Evonne interface and the core interaction mechanic (unfolding collapsed proofs by clicking), see Figure 12 in Appendix. They were presented with a short and intuitive practice proof ("A banana bread recipe containing walnuts must be labelled with an allergen note").[9] Participants were explicitly told that the system would count their clicks during the main proofs, and they should only click when necessary to understand a step.

[6] https://mt.inf.tu-dresden.de/en/research/research-projects/evonne/

[7] We had to modify our preregistration due to an error. The first version used a sensitivity analysis based on the t-test family, but for our ANOVA (as reflected in the corrected preregistration), an F-test sensitivity analysis would have been correct.

[8] https://www.prolific.com/

[9] https://www.youtube.com/watch?v=Hqdd_TbmL6o

To rule out prior domain familiarity, each participant completed two proof-understanding tasks in two made-up domains, *magical artefacts* (A) and *science fiction* (S). The order of the expert (E) and novice (N) conditions was randomized and counterbalanced across participants. More precisely, to control for order effects and domain-specific difficulty, we considered four groups: Group I: E(A), N(S); Group II: E(S), N(A); Group III: N(A), E(S); Group IV: N(S), E(A). Each participant was randomly assigned to one of these groups. This design ensures that each proof appears equally often in the expert and novice conditions, and that any effects of task fatigue or practice are distributed evenly across the experimental conditions.

Each individual task consisted of 2–3 phases. First, only participants in Condition (E) watched a short instructional video.[10] This video explains key background concepts that are directly relevant to approximately 50% of the proof steps. Next, participants were presented with the logical proof in its fully collapsed form within the Evonne interface. They had to click to unfold proof steps until they felt they understood the complete argument; see fully unfolded proofs in Figures 13, 14 in Appendix. The system logs all interactions (clicks, time) automatically. Finally, having the proof interface opened, participants completed five fill-in-the-blank questions about the proof content, using drop-down menus with three answer options each, see Figures 15, 16 in Appendix. These questions checked whether participants really understood the proof. Participants who fail to demonstrate basic understanding are excluded from the analysis, as their interaction data would not reflect a genuine comprehension process. Finally, participants rated the perceived difficulty of the comprehension questions on a Likert-like scale (from "very easy" to "very difficult").

Thus, the three measured variables are: (1) number of expansion clicks, (2) perceived difficulty (1–7 Likert-like scale) and (3) task completion time, under the instruction that participants should refer to the proof while completing the task.

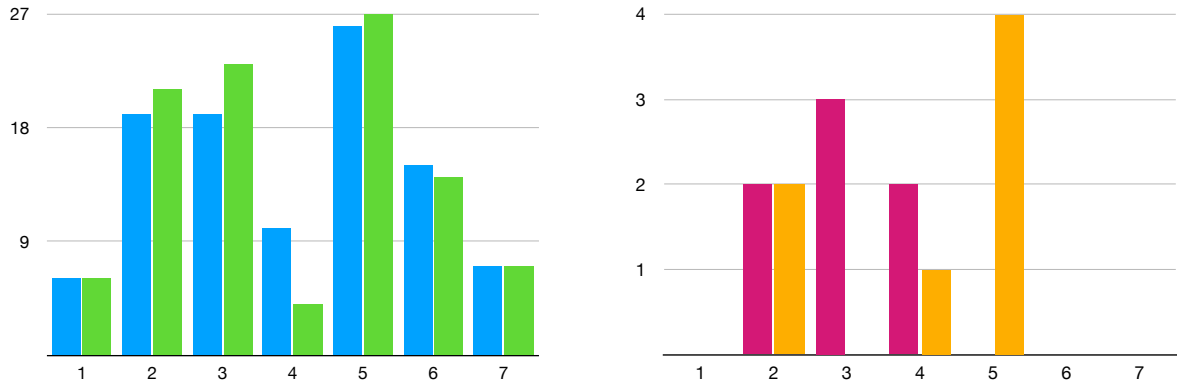
Results Analyses were calculated using R [38]. First, we constructed a correlation matrix to examine the relationships among the dependent variables separately for experts and novices. Using the Bonferroni-Holm correction to adjust for alpha error accumulation, among experts, we found significant correlations for all dependent variables: clicks and rating, $r = .21$, $p = .034$; clicks and time, $r = .28$, $p = .006$, and time and rating, $r = .29$, $p = .003$. For novices, we found a significant correlation between time and rating, $r = .29$, $p = .005$. These relationships reflect expected covariation, e.g., participants clicking more also needed more time.

To test for an overall effect, a Hotelling's T^2 test was conducted. The analysis showed no global effect of expertise on clicks, rating, or time, with $T^2 = 1.65$, $F(3, 92) = 0.54$, $p = .656$. Because there were specific hypotheses for each dependent variable, three within-subject ANOVAs were conducted. For clicks, the effect of expertise was non-significant, $F(1, 94) = 0.06$, $p = .803$, $\eta_p^2 = .001$, indicating no difference between experts and novices. Ratings also showed no significant effect, $F(1, 94) = 1.50$, $p = .224$, $\eta_p^2 = .016$. The same was true for time, where the effect of expertise was again nonsignificant with $F(1, 94) = 0.28$ and $p = .599$, $\eta_p^2 = .003$. Across all three dependent variables, the results indicate that experts and novices behaved similarly and that expertise did not influence their interaction patterns, subjective evaluations, or time spent on the tasks. Figure 3a, Figure 4a and Figure 5 demonstrate the similarity between groups.

Explorative Analysis To further evaluate the lack of effects, we conducted an explorative analysis. We constructed a dataset of 80 participants who made no mistakes in the fill-in-the-blanks text to see whether this "best" data shows any effect. A second Hotelling's T^2, computed on this subset of data, produced nearly identical results, again with no significant effect of expertise, $T^2 = 1.65$, $F(3, 77) = 0.54$, $p = .66$.Thus, both analyses consistently showed no evidence for a multivariate impact of expertise on participants' behaviour.

We furthermore conducted a paired t-test for the fill-in-the-blank comprehension task for all 95 participants. It also indicated no difference in comprehension performance among those who completed

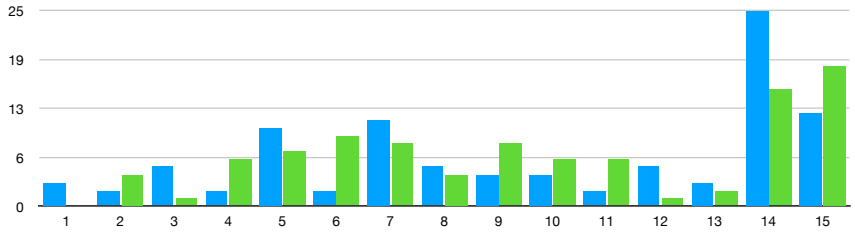
[10] https://youtu.be/H0XzfOd1Lwk or https://youtu.be/Tf9t-mtKC04.



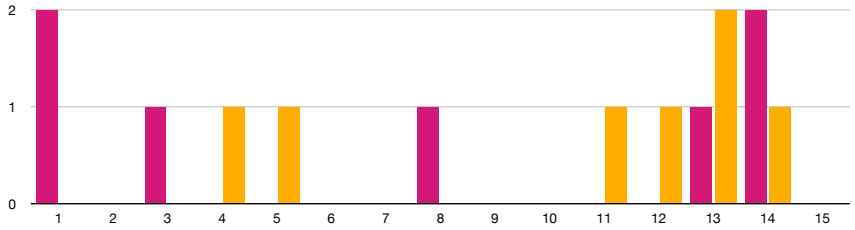
(a) General population – experts in blue, novices in green.

(b) Logicians – experts in purple, novices in orange.

Figure 3: Comparison of perceived task difficulty between experts and novices with ratings ranged from 1 ("very easy") to 7 ("very difficult").



(a) General population - experts in blue and novices in green.



(b) Logicians - experts in purple, novices in orange.

Figure 4: Comparison of the number of clicks between experts and novices. Out of 95+7 participants, 39 experts fully expanded the proof.

the task, $t(94) = 0.11$, $p = .909$. This aligns with the other analyses in showing that the manipulation of expertise (i.e., training some users to be experts) did not translate into measurable performance differences.

Discussion Several explanations could account for the lack of observed effects. One possibility is that the manipulation of expertise was not sufficiently strong [47]. Participants in the expert condition may not have internalised their role convincingly and might still have doubted their knowledge or checked back frequently. Because of this, they may not have behaved differently from novices. Another explanation is that the fictional domains themselves may be too complex or unfamiliar for participants to quickly develop genuine expertise. Achieving expert-level fluency in a fictional ontology may require far more time and immersion than was provided in the study.

Another explanation may be the affordance character of the interface: It strongly invited clicking behaviour, since buttons naturally create an impulse to interact [48]. This may have overridden more subtle expertise-related patterns, leading to similar behaviour across groups.

Additionally, the sample was recruited via Prolific, and such samples have sometimes been associated with inattentiveness, variability or engaging in unrelated tasks during testing [49, 50] that can

mask effects. In some cases, participants may have prioritised passing the attention check to secure remuneration, rather than carefully following the task instructions as intended.

Importantly, the study relied on ordinary participants rather than individuals with expertise in logic, such as those familiar with ontologies or similar formalisms. Testing real logicians could potentially yield clearer differences.

Section 5 presents additional results following the same study procedure, but with variation in the two aforementioned factors: supervision and population.

5. Study VIII – "I Already Know That": Can Explanations for Experts Be Shortened? A Qualitative Version with Logicians

For an explorative version of the same study, we also asked qualitative interview questions after the study session to collect feedback and descriptively verify whether the observed tendency in Section 4 had changed. We asked seven people (five male, two female) aged between 18 and 33 (mean = 27.7) with some knowledge of logic to complete the study with additional open-ended questions:

1. You rated Task 1 as [X] and Task 2 as [Y] in terms of difficulty. Can you walk me through what made the [easier/harder] task feel that way?
2. In the expert task (post-training), did you find yourself recognising patterns or concepts from the video? Can you give me a specific example? (If they completely expanded the proof:) Why did you expand the proof fully?
3. Did you understand one proof better than the other? And if so, why?
4. What else would have helped you understand the proofs?

Results Relating to the perceived difficulty, Figure 3b suggests a clearer separation between experts and novices compared to the Figure 3a. Expert responses (purple) are clustered around 3:"slightly easy", while novices (orange) have a sharper peak at 5:"slightly difficult". We could have improved the consistency of the answers if we had extended the training phase with a training task, to enable the participants to know what is expected of them in the main task. This would also positively affect the time measure: in Figure 5, the points are evenly scattered, partly due to the randomized order of expert and novice conditions. In fact, some participants reported that the second task felt easier, and they completed it faster because they already knew what to expect (see Figure 6).

The number of clicks on Figure 4b is spread across the entire scale, similarly to Figure 4a. From the qualitative interviews, we observed that it was possible to pass the comprehension task (by getting at least three out of five answers correctly) without looking at the proof, just following the text linguistically, which is also visible on the left sides of both charts in Figure 4. Participants also confirmed our guess that experts checking whole proofs rather than using the knowledge from the video—a pattern visible among participants on the right side of Figure 4—is due to doubts about the information they received in the video. Moreover, the study instructions were partially contradictory: completing a fill-in-the-blanks text about the proof requires expanding and studying it, but at the same time the proof should not be expanded unless necessary. Among the seven participants, some prioritised the first instruction over the second, some prioritised the second over the first, and some were finding a compromise. Therefore, measuring the number of clicks did not work as intended.

6. Recommendations for Studies on Logical Reasoning

Taking into account these and previous studies [28, 29], Table 2 presents a set of guidelines for studies on logical reasoning tasks. First, a general recommendation, which is already a common practice in fields involving user studies, is to conduct pilot studies before assessing the real sample [51]. Run a study with a few people before to check technical and content issues. Afterwards, ask follow-up questions to ensure that the instructions were as clear as possible. If possible, participants for the pilot

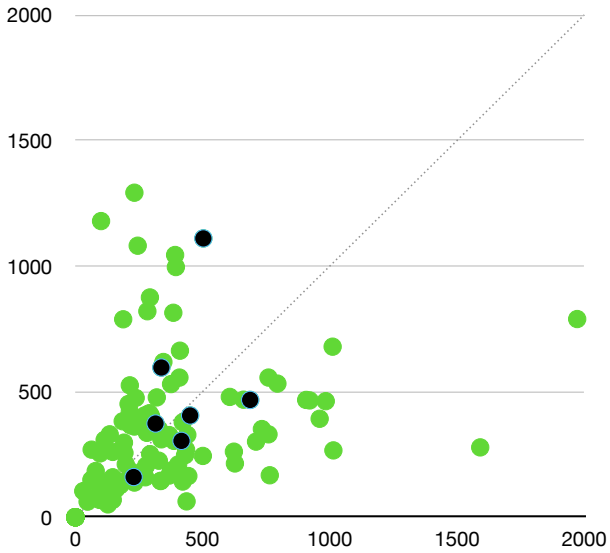


Figure 5: Comparison time taken for the expert (X-value) and novice (Y-value) tasks for each participant. The black dots are the logicians we asked additionally. There are 46 data points above the diagonal, and 56 below it.

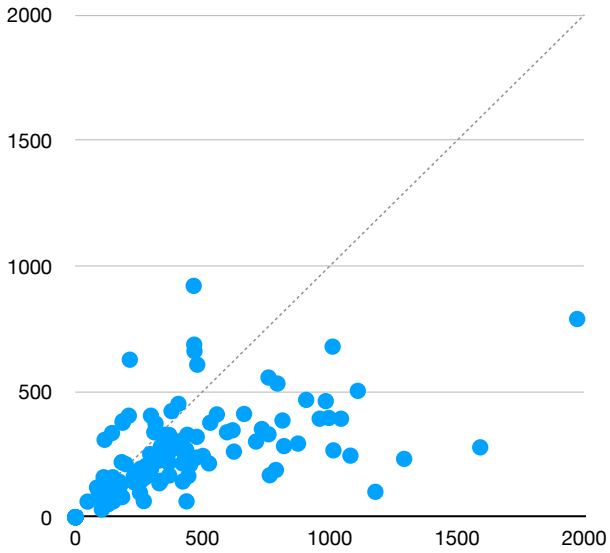


Figure 6: Comparison of the time taken for the first (X-value) and second (Y-value) tasks for each participant of Studies VII and VIII. There are 25 data points above the diagonal, and 77 below it.

study should be selected from the same population as for the main study. One also needs to avoid ceiling or floor effects by checking whether the tasks are too easy or too hard. To minimise ceiling effects, the task difficulty should be increased, for example, by adding more distractors, requiring more steps to reach a goal, or using more detailed stimuli [52, 53]. Combining easy and difficult items can also help. Ideally, participant success rates across tasks should follow a normal distribution, as this indicates an appropriate balance of task difficulty and avoids clustering at the performance extremes. Pilot studies are particularly useful for detecting such effects early. Furthermore, it is important to check whether participants properly understood the manipulation. A dedicated manipulation check could test whether, if no significant effect occurs, this is due to the manipulation failing [47].

We highly recommend conducting supervised studies rather than unsupervised questionnaires, especially using Prolific [50]. Supervised studies (with participants observed by experimenters, either in person or online) make it easier to instruct participants and ensure their full concentration. We also suggest using more repetitions for each condition, as this leads to more reliable data [32]. Another

Table 2
Guidelines for Design Choices in Formal Logic Studies, informed by Studies I–VIII. Studies I–IV were previously reported in [28]; Studies V–VIII are presented in this paper.

Aspect		Studies	Pro	Cons
Setting	Unsupervised	II, III, IV, V, VII	Fast, more participants	No direct feedback, participants may interpret instructions incorrectly
	Supervised	I, VI, VIII	Can spot problems right away	Time-consuming and monotone
Population	General	II, III, IV, V, VII	Larger pool, crowdsourcing platforms (e.g. Prolific) can be used	Struggle with formal symbols, worse data quality
	Logicians	I, VI, VIII	Higher data quality, less training required	Smaller pool, requires individual contacts for recruiting and remuneration
Method	Qualitative	VI, VIII	More detailed information about participants' preferences	More time and effort to evaluate
	Quantitative	I, II, III, IV, V, VII	State of the art in user studies, quantifiable, comparable, reliable, valid and objective data	Lack of information on participants' preferences
Domain	Expert	—	Real words, real ontology use cases	Difficult, unfamiliar words
	Common	I, V, VI	More understandable, intuitive terminology	Need to control for prior knowledge of participants
	Letters/ Nonsense	II, III, IV	Shorter, clearer logical structure	Unfamiliar for non-logicians
	Made-up	VII, VIII	No interference from prior knowledge	Effort to invent internally coherent examples
Syntax	Formal symbols	I, VI	More concise, no misinterpretations	Requires training for laypeople
	Natural text	II, III, IV, V, VII, VIII	Readable, more understandable for laypersons	Longer proofs, formula-to-text translation sacrifices precision
Training	Extensive	—	Good preparation before actual tasks, uniform baseline	Increased completion time, training fatigue
	Basic	All	Fast, reduced experimenter influence	First task may cause struggles, unpredictable behavior
Pilot Study	Extensive	—	More information on what to adjust for final study	More time, participant effort is wasted (pilot data are discarded)
	Brief	All	Less time	Less information on what to adjust for final study
Tasks	Multiple choice	II, III, IV, V, VII, VIII	Automatic evaluation, controlled responses	Enables guessing, requires precise instructions
	Open-ended	I, VI	Can spot problems right away, reveals participants' reasoning	Possible participant frustration, harder to evaluate

way to increase the knowledge gained from studies on logical reasoning is to include qualitative questions [31, 54]. Open-ended responses allow participants to describe their thoughts and preferences in more detail. Finally, another suggestion is to recruit participants who are familiar with the underlying logical formalisms. This ensures that neither the logical syntax nor the concept of a proof is confusing to the participants. On the other hand, if the study involves laypeople, we recommend using natural-language text instead of specialised logical notation.

Regarding the choice of domain for the reasoning tasks, there is no clear recommendation. Using only letters like A, B, C as concept and role names makes the logical structure clearer, but may be alienating to laypeople [28]. One the other hand, people are more comfortable with using familiar

domains about pizzas or movies, but there is a larger danger of real-world experience interfering with the logical task that is to be tested [28, 29]. Using expert domains, e.g., from biomedical ontologies, or made-up domains like in Section 4 reduces the likelihood that participants are already familiar with the domain, but increases the cognitive overhead since they have to deal with unfamiliar terms. At the extreme end, using nonsense names like "luxi" or "woal" removes any possibility of domain familiarity, but makes it much harder to remember and reason about the domain [28].

Acknowledgments

The authors thank Christian Alrabbaa for implementing the changes to Evonne for Studies VII and VIII. This work was supported by Deutsche Forschungsgemeinschaft (DFG) grant 389792660 as part of TRR 248 – CPEC, see https://perspicuous-computing.science.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] F. Baader, B. Suntisrivaraporn, Debugging SNOMED CT using axiom pinpointing in the description logic $\mathcal{EL}^+$, in: KR-MED, 2008. URL: http://ceur-ws.org/Vol-410/Paper01.pdf.

[2] S. Schlobach, R. Cornet, Non-standard reasoning services for the debugging of description logic terminologies., in: IJCAI, 2003. URL: http://ijcai.org/Proceedings/03/Papers/053.pdf.

[3] M. Horridge, Justification Based Explanation in Ontologies, Ph.D. thesis, University of Manchester, UK, 2011. URL: https://www.research.manchester.ac.uk/portal/files/54511395/FULL_TEXT.PDF.

[4] S. Schlobach, Explaining subsumption by optimal interpolation, in: JELIA, 2004. doi:10.1007/978-3-540-30227-8_35.

[5] D. Calvanese, D. Lanti, A. Ozaki, R. Peñaloza, G. Xiao, Enriching ontology-based data access with provenance, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI), ijcai.org, 2019, pp. 1616–1623. URL: https://doi.org/10.24963/ijcai.2019/224. doi:10.24963/IJCAI.2019/224.

[6] C. Bourgaux, A. Ozaki, R. Peñaloza, L. Predoiu, Provenance for the description logic elhr, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, ijcai.org, 2020, pp. 1862–1869. doi:10.24963/IJCAI.2020/258.

[7] M. Horridge, B. Parsia, U. Sattler, Justification oriented proofs in OWL, in: ISWC, 2010. doi:10.1007/978-3-642-17746-0_23.

[8] Y. Kazakov, P. Klinov, A. Stupnikov, Towards reusable explanation services in Protege, in: DL Workshop, 2017. URL: http://www.ceur-ws.org/Vol-1879/paper31.pdf.

[9] C. Alrabbaa, F. Baader, S. Borgwardt, P. Koopmann, A. Kovtunova, Finding small proofs for description logic entailments: Theory and practice, in: LPAR-23, 2020. doi:10.29007/nhpp.

[10] C. Alrabbaa, F. Baader, S. Borgwardt, P. Koopmann, A. Kovtunova, On the complexity of finding good proofs for description logic entailments, in: DL Workshop, 2020. URL: http://ceur-ws.org/Vol-2663/paper-1.pdf.

[11] C. Alrabbaa, F. Baader, S. Borgwardt, P. Koopmann, A. Kovtunova, Finding good proofs for description logic entailments using recursive quality measures, in: CADE, 2021. doi:10.1007/978-3-030-79876-5_17.

[12] C. Alrabbaa, F. Baader, S. Borgwardt, P. Koopmann, A. Kovtunova, Combining proofs for description logic and concrete domain reasoning, in: A. Fensel, A. Ozaki, D. Roman, A. Soylu (Eds.), Rules and Reasoning - 7th International Joint Conference, RuleML+RR 2023, Oslo, Norway, September 18-20, 2023, Proceedings, Lecture Notes in Computer Science, Springer, 2023, pp. 54–69. doi:10.1007/978-3-031-45072-3_4.

[13] P. Koopmann, W. Del-Pinto, S. Tourret, R. A. Schmidt, Signature-based abduction for expressive description logics, in: D. Calvanese, E. Erdem, M. Thielscher (Eds.), Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning, KR 2020, Rhodes, Greece, September 12-18, 2020, 2020, pp. 592–602. URL: https://doi.org/10.24963/kr.2020/59. doi:10.24963/KR.2020/59.

[14] F. Haifani, P. Koopmann, S. Tourret, C. Weidenbach, Connection-minimal abduction in *EL* via translation to FOL, in: J. Blanchette, L. Kovács, D. Pattinson (Eds.), Automated Reasoning - 11th International Joint Conference, IJCAR, Lecture Notes in Computer Science, Springer, 2022, pp. 188–207. doi:10.1007/978-3-031-10769-6_12.

[15] C. Elsenbroich, O. Kutz, U. Sattler, A case for abductive reasoning over ontologies, in: B. C. Grau, P. Hitzler, C. Shankey, E. Wallace (Eds.), Proceedings of the OWLED*06 Workshop on OWL: Experiences and Directions, CEUR Workshop Proceedings, CEUR-WS.org, 2006. URL: https://ceur-ws.org/Vol-216/submission_25.pdf.

[16] M. Homola, J. Pukancová, J. Boborová, I. Balintová, Merge, explain, iterate: A combination of MHS and MXP in an ABox abduction solver, in: S. A. Gaggl, M. V. Martinez, M. Ortiz (Eds.), Logics in Artificial Intelligence - 18th European Conference (JELIA), Lecture Notes in Computer Science, Springer, 2023, pp. 338–352. doi:10.1007/978-3-031-43619-2_24.

[17] J. Kloc, J. Boborová, M. Homola, J. Pukancová, CATS solver: The rise of hybrid abduction algorithms, in: L. Tendera, Y. Ibáñez-García, P. Koopmann (Eds.), Proceedings of the 38th International Workshop on Description Logics (DL), CEUR Workshop Proceedings, CEUR-WS.org, 2025. URL: https://ceur-ws.org/Vol-4091/paper34.pdf.

[18] J. Bauer, U. Sattler, B. Parsia, Explaining by example: Model exploration for ontology comprehension, in: B. C. Grau, I. Horrocks, B. Motik, U. Sattler (Eds.), Proceedings of the 22nd International Workshop on Description Logics (DL), CEUR Workshop Proceedings, CEUR-WS.org, 2009. URL: https://ceur-ws.org/Vol-477/paper_37.pdf.

[19] C. Alrabbaa, W. Hieke, Explaining non-entailment by model transformation for the description logic el, in: Proceedings of the 11th International Joint Conference on Knowledge Graphs, IJCKG'22, ACM, 2023, p. 1–9. doi:10.1145/3579051.3579060.

[20] C. Alrabbaa, S. Borgwardt, T. Friese, P. Koopmann, M. Kotlov, Why not? explaining missing entailments with evee, in: O. Kutz, C. Lutz, A. Ozaki (Eds.), Proceedings of the 36th International Workshop on Description Logics (DL 2023) co-located with the 20th International Conference on Principles of Knowledge Representation and Reasoning and the 21st International Workshop on Non-Monotonic Reasoning (KR 2023 and NMR 2023)., Rhodes, Greece, September 2-4, 2023, CEUR Workshop Proceedings, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3515/paper-1.pdf.

[21] M. R. G. Schiller, B. Glimm, Towards explicative inference for OWL, in: DL Workshop, 2013. URL: http://ceur-ws.org/Vol-1014/paper_36.pdf.

[22] F. Engström, A. R. Nizamani, C. Strannegård, Generating comprehensible explanations in description logic, in: DL Workshop, 2014. URL: http://ceur-ws.org/Vol-1193/paper_17.pdf.

[23] T. A. T. Nguyen, R. Power, P. Piwek, S. Williams, Predicting the understandability of OWL inferences, in: ESWC, 2013. doi:10.1007/978-3-642-38288-8_8.

[24] C. Alrabbaa, F. Baader, R. Dachselt, A. Kovtunova, J. Méndez, The concrete Evonne: Visualization meets concrete domain reasoning, in: R. Thiemann, C. Weidenbach (Eds.), Frontiers of Combining Systems - 15th International Symposium (FroCoS), Lecture Notes in Computer Science, Springer, 2025, pp. 3–21. doi:10.1007/978-3-032-04167-8_1.

[25] M. Horridge, B. Parsia, U. Sattler, Lemmas for justifications in OWL, in: Proceedings of the 22nd International Workshop on Description Logics (DL 2009), Oxford, UK, July 27-30, 2009, 2009. URL: http://ceur-ws.org/Vol-477/paper_24.pdf.

[26] S. Borgwardt, A. Hirsch, A. Kovtunova, F. Wiehr, In the Eye of the Beholder: Which Proofs are Best?, in: DL Workshop, 2020. URL: http://ceur-ws.org/Vol-2663/paper-6.pdf.

[27] C. Alrabbaa, S. Borgwardt, N. Knieriemen, A. Kovtunova, A. M. Rothermel, F. Wiehr, In the hand of the beholder: Comparing interactive proof visualizations, in: DL Workshop, 2021. URL: http://ceur-ws.org/Vol-2954/paper-2.pdf.

[28] C. Alrabbaa, S. Borgwardt, A. Hirsch, N. Knieriemen, A. Kovtunova, A. M. Rothermel, F. Wiehr, In the head of the beholder: Comparing different proof representations, in: G. Governatori, A. Turhan (Eds.), Rules and Reasoning - 6th International Joint Conference on Rules and Reasoning (RuleML+RR), Lecture Notes in Computer Science, Springer, 2022, pp. 211–226. doi:10.1007/978-3-031-21541-4_14.
[29] C. Alrabbaa, S. Borgwardt, T. Friese, A. Hirsch, N. Knieriemen, P. Koopmann, A. Kovtunova, A. Krüger, A. Popovic, I. S. R. Siahaan, Explaining reasoning results for OWL ontologies with evee, in: P. Marquis, M. Ortiz, M. Pagnucco (Eds.), Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024, 2024. doi:10.24963/KR.2024/67.
[30] M. R. G. Schiller, F. Schiller, B. Glimm, Testing the adequacy of automated explanations of EL subsumptions, in: DL Workshop, 2017. URL: http://ceur-ws.org/Vol-1879/paper43.pdf.
[31] J. W. Creswell, J. D. Creswell, Research design: Qualitative, quantitative, and mixed methods approaches, Sage publications, 2017.
[32] R. Rosenthal, R. L. Rosnow, Essentials of behavioral research: Methods and data analysis, 2008.
[33] S. Borgwardt, A. Hirsch, N. Knieriemen, A. Kovtunova, In the heart of the beholder: User-tailored explanations for description logics. User study supplementary materials, 2026. doi:10.5281/zenodo.19728229.
[34] C. Alrabbaa, S. Borgwardt, A. Hirsch, N. Knieriemen, A. Kovtunova, A. Krüger, Comparing the comprehensibility of explanations for logical non-consequences, 2023. URL: osf.io/4bjmx.
[35] E. Erdfelder, F. Faul, A. Buchner, Gpower: A general power analysis program, Behavior Research Methods, Instruments, & Computers 28 (1996) 1–11.
[36] F. Faul, E. Erdfelder, A.-G. Lang, A. Buchner, G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences, Behavior Research Methods 39 (2007) 175–191. doi:10.3758/bf03193146.
[37] F. Faul, E. Erdfelder, A. Buchner, A.-G. Lang, Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses, Behavior Research Methods 41 (2009) 1149–1160. doi:10.3758/brm.41.4.1149.
[38] RStudio Team, RStudio: Integrated Development Environment for R, RStudio, PBC, Boston, MA, 2021. URL: http://www.rstudio.com/.
[39] R. Bono, Non-normal data: Is ANOVA still a valid option?, Psicothema (2017). doi:10.7334/psicothema2016.383.
[40] E. Schmider, M. Ziegler, E. Danay, L. Beyer, M. Bühner, Is it really robust?, Methodology (2010). doi:10.1027/1614-2241/a000016.
[41] M. A. Musen, The protégé project: a look back and a look forward, AI Matters 1 (2015) 4–12. doi:10.1145/2757001.2757003.
[42] C. Alrabbaa, S. Borgwardt, T. Friese, A. Hirsch, N. Knieriemen, P. Koopmann, A. Kovtunova, A. Krüger, A. Popovic, I. Siahaan, Explaining reasoning results for owl ontologies with evee - resources, 2024. doi:10.5281/zenodo.11198049.
[43] C. Rivas, Coding and analysing qualitative data, in: Researching society and culture, volume 3, SAGE, 2012, p. 367–392.
[44] C. Alrabbaa, F. Baader, S. Borgwardt, R. Dachselt, P. Koopmann, J. Méndez, Evonne: Interactive proof visualization for description logics (system description), in: IJCAR, 2022. doi:10.1007/978-3-031-10769-6_16.
[45] A. Hirsch, N. Knieriemen, A. Kovtunova, A3 study 6 - "i already know that": Can explanations for experts be shortened?, 2025. doi:10.17605/OSF.IO/D7B92.
[46] M. T. Chi, Laboratory methods for assessing experts' and novices' knowledge, The Cambridge handbook of expertise and expert performance (2006) 167–184.
[47] B. C. Perdue, J. O. Summers, Checking the success of manipulations in marketing experiments, Journal of marketing research 23 (1986) 317–326.
[48] R. Hartson, Cognitive, physical, sensory, and functional affordances in interaction design, Behaviour & information technology 22 (2003) 315–338.

[49] A. C. Drody, E. J. Pereira, D. Smilek, A desire for distraction: uncovering the rates of media multitasking during online research studies, Scientific Reports 13 (2023) 781. doi:10.1038/s41598-023-27606-3.
[50] E. Peer, L. Brandimarte, S. Samat, A. Acquisti, Beyond the turk: Alternative platforms for crowdsourcing behavioral research, Journal of experimental social psychology 70 (2017) 153–163.
[51] E. Van Teijlingen, V. Hundley, The importance of pilot studies, Social research update (2001) 1–4.
[52] L. Crocker, J. Algina, Introduction to classical and modern test theory., ERIC, 1986.
[53] S. E. Embretson, S. P. Reise, Item response theory for psychologists, Psychology Press, 2013.
[54] G. Guest, K. M. MacQueen, E. E. Namey, Applied thematic analysis, Sage Publications, 2011.

A. User Study Trainings and Task Examples

This section contains supplementary material for the paper.

Trainingsphase

Bitte lesen Sie sich alles gründlich durch und versuchen Sie, es zu verstehen. Sie können jederzeit zu dieser Seite zurückkehren.

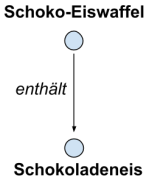
Im Rahmen der Umfrage sehen Sie eine Liste von Aussagen wie die folgenden:

(1) Jede **Schoko-Eiswaffel** *enthält* **Schokoladeneis**.
(2) Eine **reine Schoko-Eiswaffel** ist alles, was eine **Schoko-Eiswaffel** ist und nur **Schokoladeneis** *enthält*.
(3) **Schlagsahne** ist kein **Schokoladeneis**.

Fettschrift wird für Begriffe (z. B. **reine Schoko-Eiswaffel**, **Schokoladeneis**) und Schreibschrift für Relationen (z. B. *enthält*) verwendet.

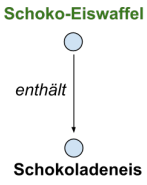
Es gibt zwei Arten von Aussagen, denen Sie begegnen werden:
- Definitionen/Äquivalenzen, formuliert mittels "alles, was", z. B. (2).
- Subsumptionen/Implikationen mit "jede" oder "ist", z. B. (1) und (3).

Wir zeigen Modelle von Aussagen als gerichtete Graphen. Ein Modell von (1) ist zum Beispiel das folgende:

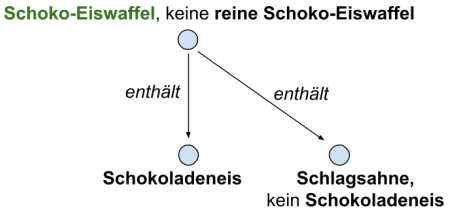


Die Knoten im Graph stellen Elemente des Modells dar. Die Bezeichnungen eines Knotens stehen für die Begriffe, zu denen das entsprechende Element gehört. In diesem Beispiel haben wir zwei Elemente. Eines ist eine **Schoko-Eiswaffel**, und das andere ein **Schokoladeneis**. Eine Kante zwischen zwei Knoten stellt eine Relation zwischen den entsprechenden Elementen dar. Im Beispiel ist die **Schoko-Eiswaffel** mit dem **Schokoladeneis** über die Relation *enthält* verbunden.

Wir heben besondere Begriffe im Graphen in Grün hervor:



Um hervorzuheben, dass ein Element nicht zu einem bestimmten Begriff gehört, schreiben wir es mit dem Schlüsselwort "kein(e)" oder "nicht".



In diesem Graph heißt das nicht, dass die **Schoko-Eiswaffel** kein **Schokoladeneis** *enthält*, sondern nur, dass die **Schoko-Eiswaffel** etwas *enthält* (nämlich **Schlagsahne**), das kein **Schokoladeneis** ist.

Der obige Graph erfüllt alle drei Aussagen (1)-(3). Folgende Aussagen folgen also *nicht* aus (1)-(3), da der Graph ein Gegenbeispiel dazu zeigt:

Jede **Schoko-Eiswaffel** *enthält* nur **Schokoladeneis**.
Jede **Schoko-Eiswaffel** ist eine **reine Schoko-Eiswaffel**.

Wir verwenden auch die Farbe Blau, um besondere Teile eines Graphen hervorzuheben.

Figure 7: Training for the counterexample group.

Beispiel 1

Unten sehen Sie eine Aussage, die nicht aus der Liste folgt (die dazu relevanten Aussagen sind hervorgehoben). Darunter finden Sie eine Erklärung dazu. Bitte lesen Sie sich alles gründlich durch und versuchen Sie nachzuvollziehen, warum die Aussage nicht folgt.

Jeder **Jagdfilm** ist ein **Film**. Jeder **Jagdfilm** *zeigt* einen **Hund**. Jeder **Jagdfilm** *zeigt* einen **Reiter**. Jeder **Jagdfilm** *zeigt* eine **Tötung**. Jeder **Westernfilm** ist ein **Film**. Jeder **Westernfilm** *zeigt* einen **Reiter**. Jeder **Westernfilm** *zeigt* ein **Pferd**. Jeder **Westernfilm** *zeigt* einen **Kampf**. Jeder **Bauernhoffilm** ist ein **Film**. Jeder **Bauernhoffilm** *zeigt* einen **Hund**. Jeder **Bauernhoffilm** *zeigt* einen **Reiter**. Jeder **Film** *zeigt* ein **Handlungselement**. **FSK-16-Filme** sind alle **Filme**, die **Gewalt** *zeigen*. **Tierfilme** sind alle **Filme**, die ein **Tier** *zeigen*.	**FSK-12-Filme** sind alle **Filme**, die keine **Gewalt** *zeigen*. **Kinderfilme** sind alle **Filme**, die nur **Freundliches** *zeigen*. **Dramen** sind alles, was etwas *zeigt*, das sowohl **freundlich** als auch **Gewalt** ist. Jeder **Reiter** ist **freundlich**. Jeder **Hund** ist ein **Tier**. Jedes **Tier** ist ein **Lebewesen**. Jedes **Tier** ist ein **Handlungselement**. Jeder **Kampf** ist **Gewalt**. Jede **Gewalt** ist ein **Handlungselement**. Eine **Zeichentrickfigur** ist ein **freundliches Lebewesen**. Jedes **Tier** *ruft* **Glück** *hervor*. **Glück** ist ein **Gefühl**. Jede **Gewalt** *ruft* **Angst** *hervor*. **Angst** ist ein **Gefühl**.

Die Aussage
Jeder **Jagdfilm** ist ein **FSK-16-Film**.

folgt nicht aus den obigen Aussagen.

Unten sehen Sie eine automatisch generierte Erklärung dazu:

Die obige Aussage folgt nicht, weil keine der folgenden Aussagen zutrifft:

- Jeder **Hund** ist **Gewalt**.
- Jeder **Reiter** ist **Gewalt**.
- Jede **Tötung** ist **Gewalt**.
- Jedes **Tier** ist **Gewalt**.
- Alles **Freundliche** ist **Gewalt**.
- Jedes **Handlungselement** ist **Gewalt**.
- Jeder **Hund** ist ein **Kampf**.
- Jeder **Reiter** ist ein **Kampf**.
- Jede **Tötung** ist ein **Kampf**.
- Jedes **Tier** ist ein **Kampf**.
- Alles **Freundliche** ist ein **Kampf**.
- Jedes **Handlungselement** ist ein **Kampf**.
- Jeder **Jagdfilm** *zeigt* **Gewalt**.
- Jeder **Jagdfilm** *zeigt* einen **Kampf**.

*Wie gut haben Sie verstanden, dass die obige Aussage nicht folgt (1 = gar nicht, 5 = sehr gut)?

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

*Wie hilfreich fanden Sie dazu die Erklärung (1 = gar nicht hilfreich, 5 = sehr hilfreich)?

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 8: Example of an abductive explanation of a missing entailment in the movie domain.

Beispiel 2

Unten sehen Sie eine Aussage, die nicht aus der Liste folgt (die dazu relevanten Aussagen sind hervorgehoben). Darunter finden Sie eine Erklärung dazu. Bitte lesen Sie sich alles gründlich durch und versuchen Sie nachzuvollziehen, warum die Aussage nicht folgt.

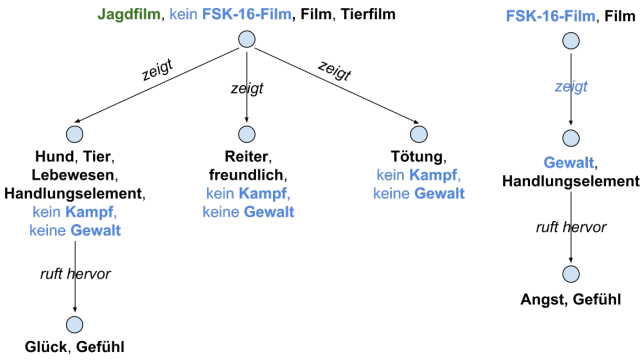
Jeder **Jagdfilm** ist ein **Film**. Jeder **Jagdfilm** *zeigt* einen **Hund**. Jeder **Jagdfilm** *zeigt* einen **Reiter**. Jeder **Jagdfilm** *zeigt* eine **Tötung**. Jeder **Westernfilm** ist ein **Film**. Jeder **Westernfilm** *zeigt* einen **Reiter**. Jeder **Westernfilm** *zeigt* ein **Pferd**. Jeder **Westernfilm** *zeigt* einen **Kampf**. Jeder **Bauernhoffilm** ist ein **Film**. Jeder **Bauernhoffilm** *zeigt* einen **Hund**. Jeder **Bauernhoffilm** *zeigt* einen **Reiter**. Jeder **Film** *zeigt* ein **Handlungselement**. **FSK-16-Filme** sind alle **Filme**, die **Gewalt** *zeigen*. **Tierfilme** sind alle **Filme**, die ein **Tier** *zeigen*.	**FSK-12-Filme** sind alle **Filme**, die keine **Gewalt** *zeigen*. **Kinderfilme** sind alle **Filme**, die nur **Freundliches** *zeigen*. **Dramen** sind alles, was etwas *zeigt*, das sowohl **freundlich** als auch **Gewalt** ist. Jeder **Reiter** ist **freundlich**. Jeder **Hund** ist ein **Tier**. Jedes **Tier** ist ein **Lebewesen**. Jedes **Tier** ist ein **Handlungselement**. Jeder **Kampf** ist **Gewalt**. Jede **Gewalt** ist ein **Handlungselement**. Eine **Zeichentrickfigur** ist ein **freundliches Lebewesen**. Jedes **Tier** *ruft* **Glück** *hervor*. **Glück** ist ein **Gefühl**. Jede **Gewalt** *ruft* **Angst** *hervor*. **Angst** ist ein **Gefühl**.

Die Aussage
Jeder **Jagdfilm** ist ein **FSK-16-Film**.

folgt nicht aus den obigen Aussagen.

Unten sehen Sie eine automatisch generierte Erklärung dazu:

Die obige Aussage folgt nicht, weil es folgendes Modell gibt:



*Wie gut haben Sie verstanden, dass die obige Aussage nicht folgt (1 = gar nicht, 5 = sehr gut)?

1 2 3 4 5

*Wie hilfreich fanden Sie dazu die Erklärung (1 = gar nicht hilfreich, 5 = sehr hilfreich)?

1 2 3 4 5

Figure 9: Example of a counterexample for a missing entailment in the movie domain.

Aufgabe 2

Unten sehen Sie erneut die Liste von Aussagen. Bitte beantworten Sie darunter, ob die angegebenen Aussagen aus den oberen folgen.

Jeder **Jagdfilm** ist ein **Film**. Jeder **Jagdfilm** *zeigt* einen **Hund**. Jeder **Jagdfilm** *zeigt* einen **Reiter**. Jeder **Jagdfilm** *zeigt* eine **Tötung**. Jeder **Westernfilm** ist ein **Film**. Jeder **Westernfilm** *zeigt* einen **Reiter**. Jeder **Westernfilm** *zeigt* ein **Pferd**. Jeder **Westernfilm** *zeigt* einen **Kampf**. Jeder **Bauernhoffilm** ist ein **Film**. Jeder **Bauernhoffilm** *zeigt* einen **Hund**. Jeder **Bauernhoffilm** *zeigt* einen **Reiter**. Jeder **Film** *zeigt* ein **Handlungselement**. **FSK-16-Filme** sind alle **Filme**, die **Gewalt** *zeigen*. **Tierfilme** sind alle **Filme**, die ein **Tier** *zeigen*.	**FSK-12-Filme** sind alle **Filme**, die keine **Gewalt** *zeigen*. **Kinderfilme** sind alle **Filme**, die nur **Freundliches** *zeigen*. **Dramen** sind alles, was etwas *zeigt*, das sowohl **freundlich** als auch **Gewalt** ist. Jeder **Reiter** ist **freundlich**. Jeder **Hund** ist ein **Tier**. Jedes **Tier** ist ein **Lebewesen**. Jedes **Tier** ist ein **Handlungselement**. Jeder **Kampf** ist **Gewalt**. Jede **Gewalt** ist ein **Handlungselement**. Eine **Zeichentrickfigur** ist ein **freundliches Lebewesen**. Jedes **Tier** *ruft* **Glück** *hervor*. **Glück** ist ein **Gefühl**. Jede **Gewalt** *ruft* **Angst** *hervor*. **Angst** ist ein **Gefühl**.

*Welche der Aussagen folgen aus den obigen?

	Ja	Nein	Ich weiß es nicht
Jeder **Westernfilm** *zeigt* einen **Hund**.	○	○	○
Jedes **Pferd** *ruft* ein **Gefühl** *hervor*.	○	○	○
Jeder **Hund** *ruft* ein **Gefühl** *hervor*.	○	○	○
Jedes **Lebewesen** *ruft* ein **Gefühl** *hervor*.	○	○	○
Jeder **FSK-16-Film** *zeigt* einen **Kampf**.	○	○	○
Jeder **FSK-16-Film** *zeigt* etwas, das **Glück** *hervorruft*.	○	○	○
Jeder **Westernfilm** *zeigt* **Gewalt**.	○	○	○
Jeder **Westernfilm** *zeigt* ein **Handlungselement**.	○	○	○
Jeder **Westernfilm** ist ein **Tierfilm**.	○	○	○
Jedes **Drama** ist ein **Tierfilm**.	○	○	○

*Wie hilfreich fanden das Beispiel auf der vorherigen Seite zum Beantworten dieser Fragen (1 = gar nicht hilfreich, 5 = sehr hilfreich)?

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 10: Task example in the movie domain.

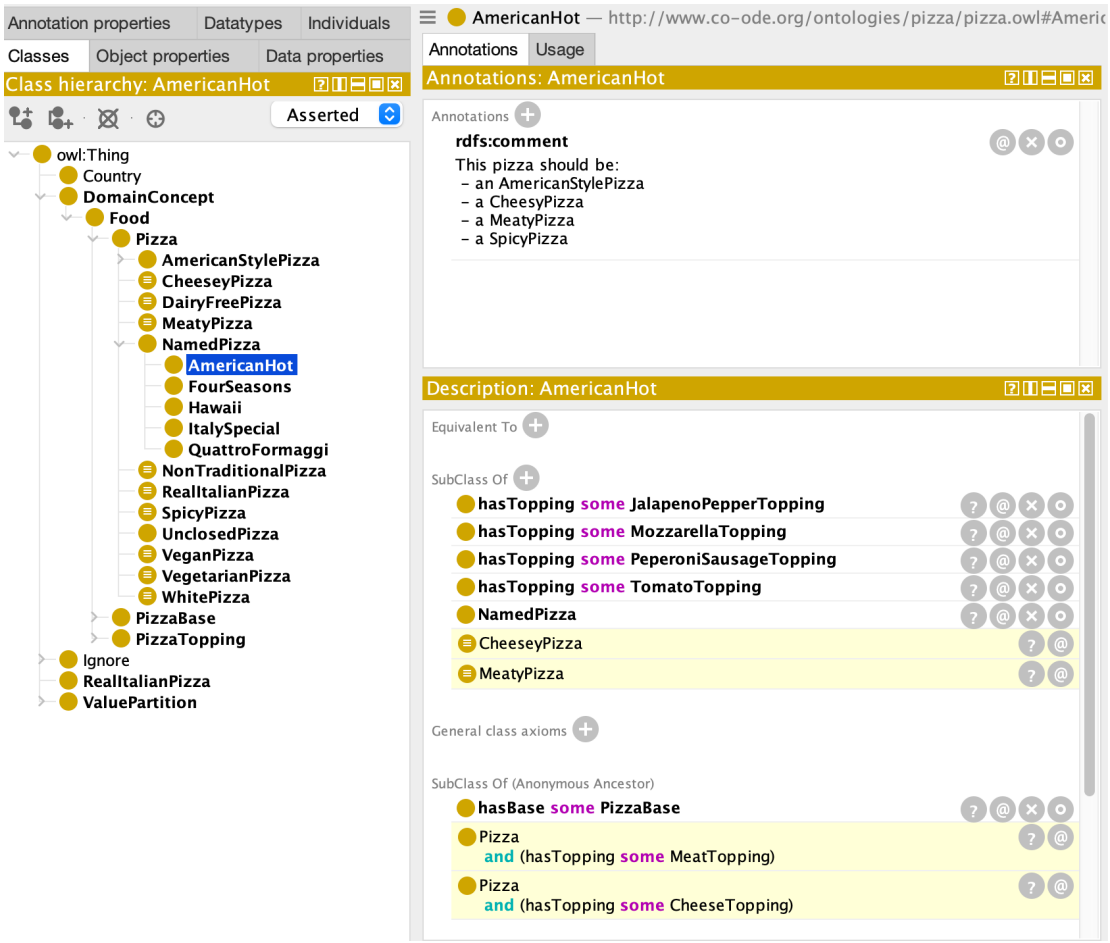


Figure 11: Pizza task example in Protégé. The AmericanHot pizza must be a SpicyPizza (see the comment) due to JalapenoPepperTopping; however, after running the reasoner, this subclass relation is not derived.

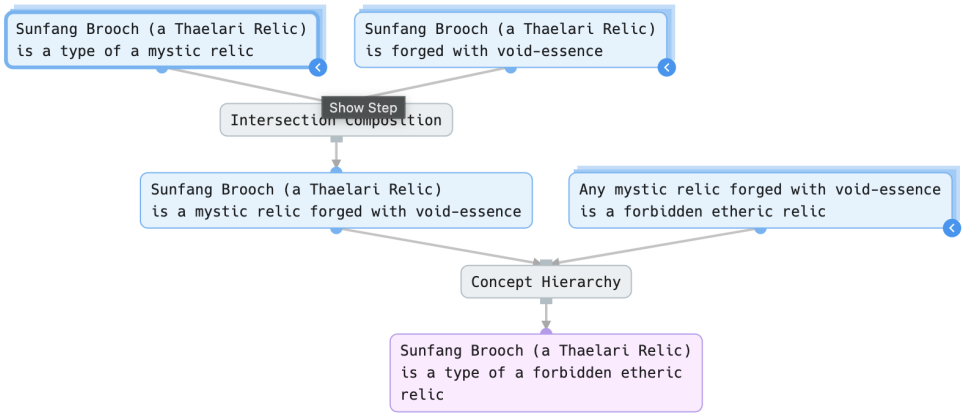


Figure 12: Two step proof unfolding in Evonne. The proof displays step by step why the final conclusion, "Sunfang Brooch (A Thaelari Relic) is a type of a forbidden etheric relic" in a purple box, holds. By clicking the "less than" symbol in the lower right corner of boxes, you can start unfolding the proof tree. The first level shows that the conclusion comes from two statements, where the left one in turn comes from another two statements on the second level.

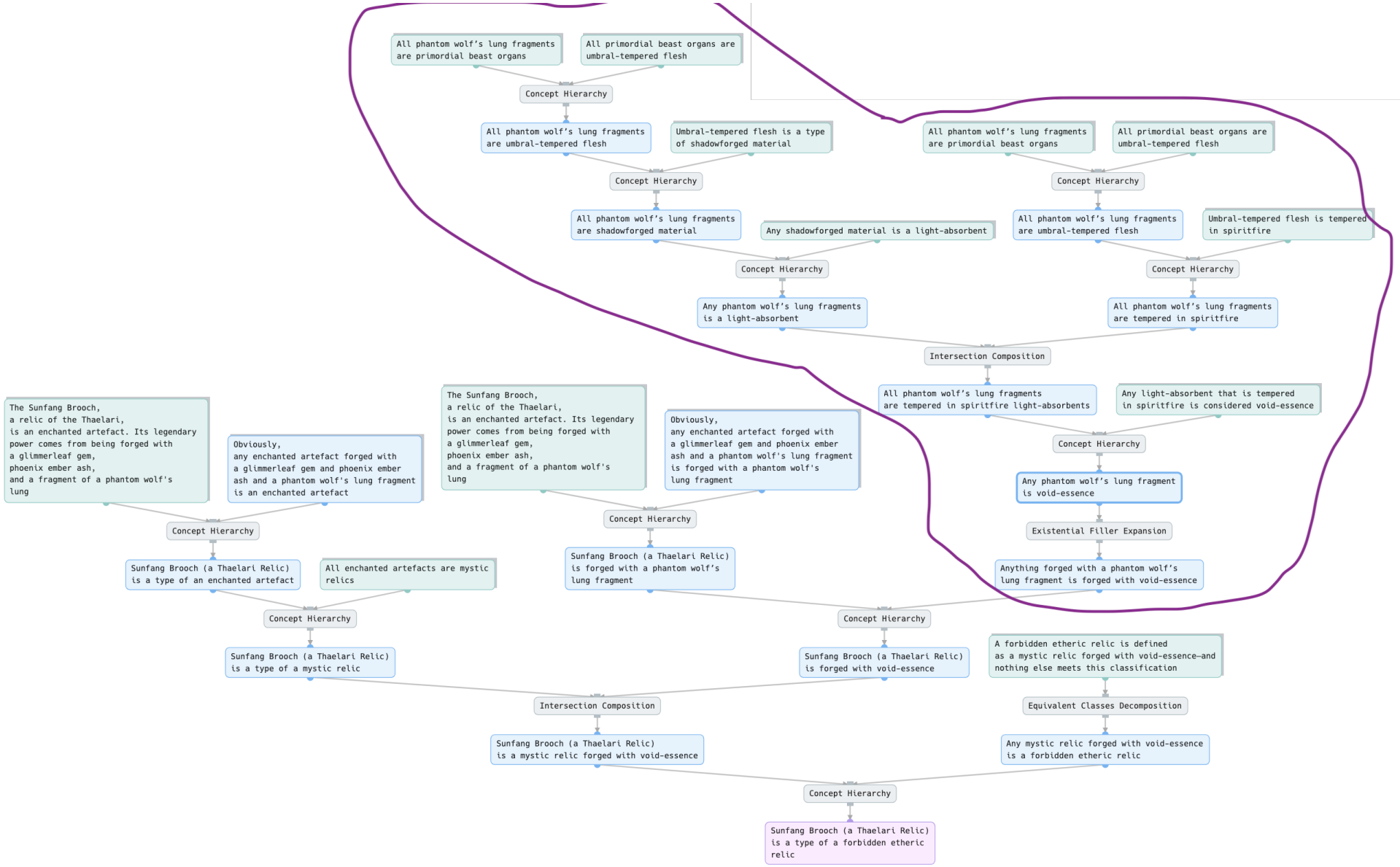


Figure 13: The fully expanded artefact proof in Evonne. The highlighted part is explained in the video. This proof is available at http://141.76.67.139:7007/proof?id=Artifact subject to server stability.

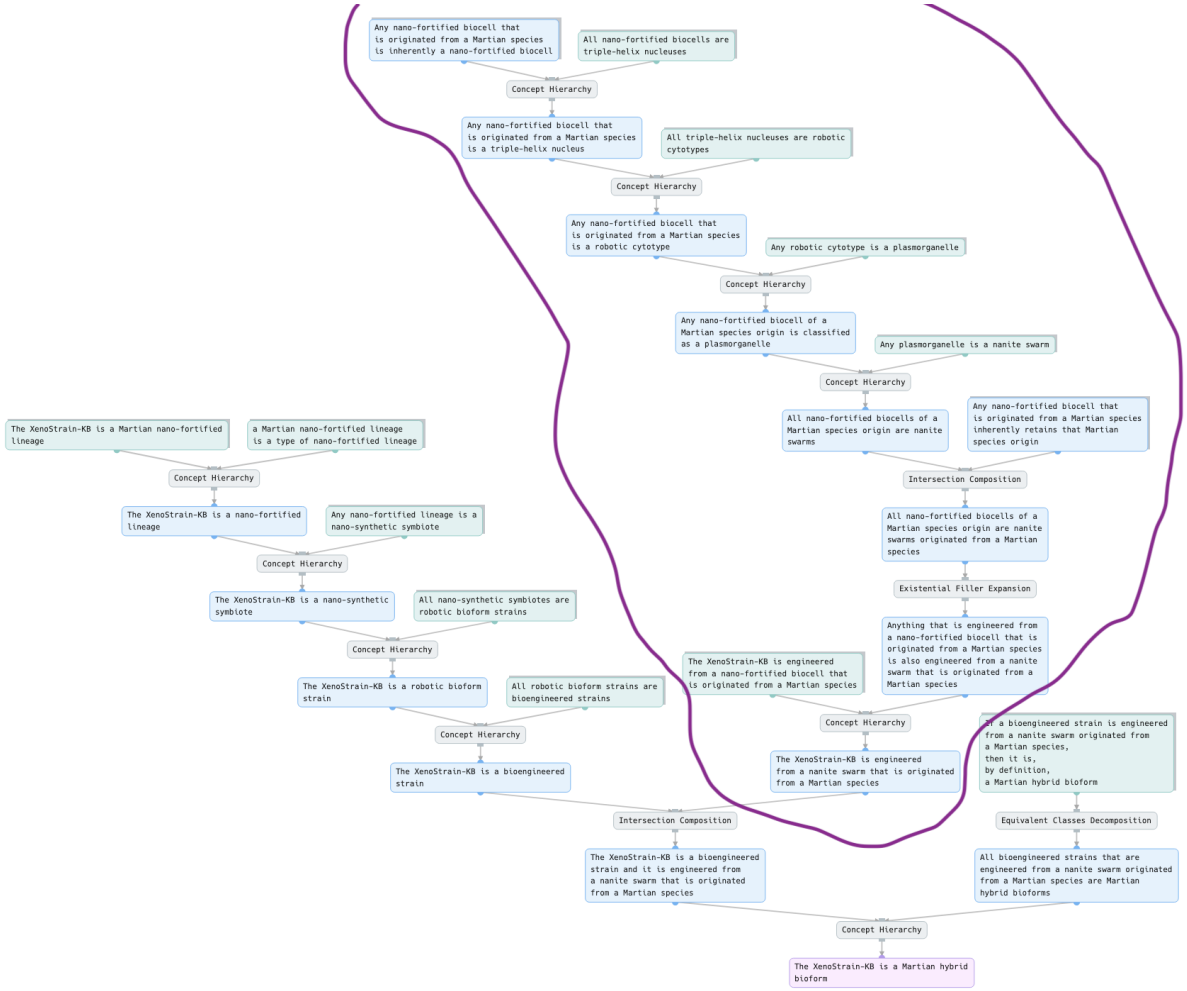


Figure 14: The fully expanded sci-fi proof in Evonne. The highlighted part is explained in the video. This proof is available at http://141.76.67.139:7007/proof?id=SciFi subject to server stability.

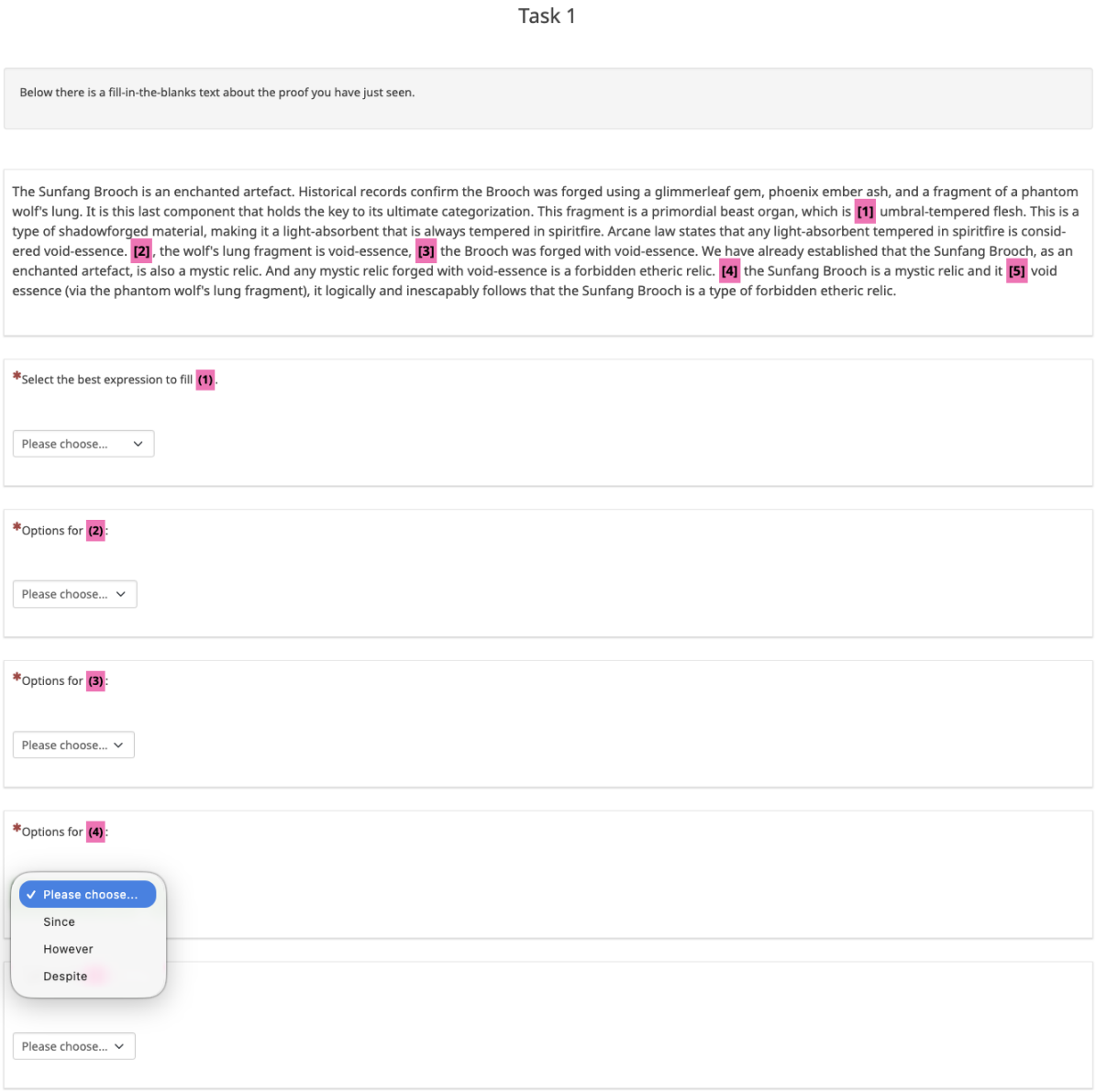


Figure 15: The after-proof task contains a fill-the-gaps (cloze) text that summarizes the artefact proof. Each of the five gaps has a 3-option multiple-choice selection.

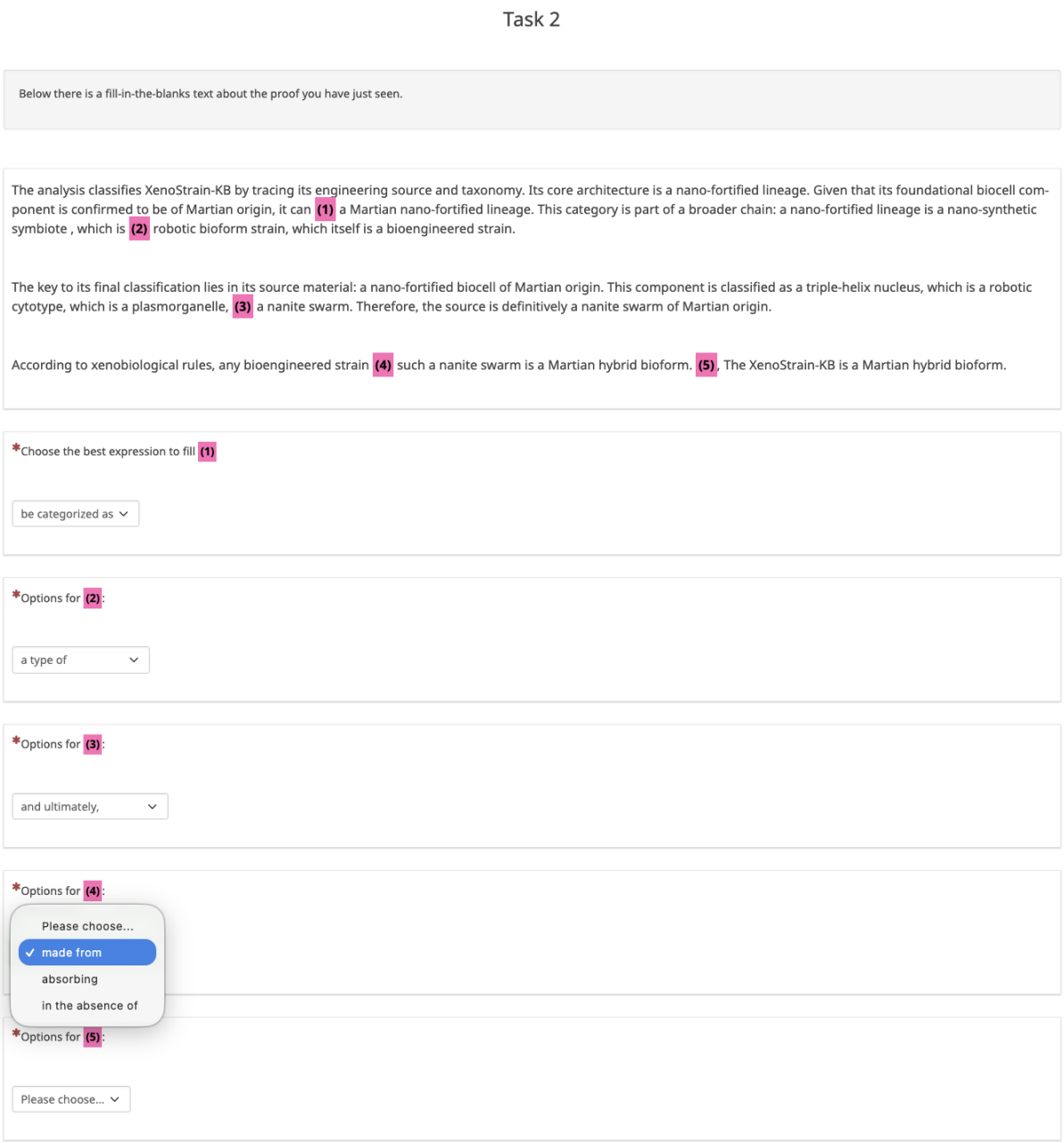


Figure 16: The after-proof task contains a fill-the-gaps (cloze) text that summarizes the sci-fi proof. Each of the five gaps has a 3-option multiple-choice selection.