Learning Comes from Experience: The Effects on Human Learning and Performance of a Virtual Assistant for Design Space Exploration

Antoni Viros i Martin and Daniel Selva

Abstract This paper studies the effects on learning and performance for a human using a virtual assistant to perform a design space exploration task—design a satellite constellation for Earth Observation. We conducted a study at Texas A&M University with N = 18 STEM students, who were asked to use two versions of an assistant to perform the task. One version acted as an "Assistant", answering questions from the user, while the other version acted as a "Peer", giving its opinion and recommendations to the user. Subjects completed surveys on usability and trust in automation, two learning tests, and a subjective survey on their learning and on the virtual assistant. The more subjects used and interacted with the assistant, the more they learned about the problem. While these findings are limited to a particular application and student population, they provide some evidence that virtual assistants can improve learning in design space exploration.

1 Introduction

Virtual Assistants (VAs), also known as cognitive assistants, have surged in popularity in the last 8 years, after the appearance of consumer products from Amazon, Microsoft, Google, and Apple. The catalyst of the modern vision of what a VA is and can do can be traced back to CALO (Cognitive Assistant that Learns and Organizes) [1], a project from DARPA to integrate existing AI technologies to create a cognitive assistant, which was redefined to mean a software agent that responds to human commands, questions in natural language, and performs tasks based on that input. Siri, the commercial VA from Apple, is a spin-off of CALO.

There is a long history of using a variety of AI systems for the design of complex systems. They have been used to represent and generate design alternatives [2], to search the solution space [3], to evaluate alternatives [4], and to provide interactive visualizations for the designer [5, 6].

A. Viros i Martin (✉) · D. Selva
Texas A&M University, College Station, USA
e-mail: aviros@tamu.edu

J. S. Gero (ed.), *Design Computing and Cognition'20*,
https://doi.org/10.1007/978-3-030-90625-2_39

In most of these systems, either the AI is a tool to the human designer [7] or the human is an input to an automated tool [8], but it is rare to find cases where collaboration between human and machine is emphasized. In contrast, current research in human—machine interaction suggests that a more collaborative approach may increase performance [9].

In this context, we started developing Daphne [10] three years ago, with the purpose of bringing the usability and cognitive unloading abilities of general VAs to the early design of complex systems, specifically Earth observation satellites. Daphne is centered on improving design space exploration, which is a vital part of the early design of complex systems.

Feedback gathered during exit interviews in a prior study with practitioners emphasized the importance of being able to justify the decisions from such studies [10]. This means the slightly negative trend for learning we observed in the experiment in [10] is worrisome, even though the method we used to measure learning in that experiment may not take into account some kinds of knowledge a test subject might gain while using Daphne. Even though we showed that performance does increase when using Daphne at its full capability, this means nothing if designers cannot justify the outputs of Daphne to stakeholders. This realization prompted development of explanation strategies for Daphne [11] and the study described in this paper.

In this study, we evaluate the response from users to different parts of Daphne, and we evaluate how learning and understanding of the problem changes based on the usage of Daphne, in hope of extracting conclusions on how to design VAs for engineering design problems that can both help improve performance and learning.

We define learning in the context of design exploration as improving the understanding of the structure of the design space, e.g., the trade-offs between different design criteria, the sensitivities of design criteria to design decisions, or the existence of families of similar designs with similar performance. Currently, there are no agreed upon metrics to measure human learning in design. Bang and Selva, inspired by Bloom's taxonomy of learning [12], proposed that these metrics should encompass different cognitive processes such as remembering information, understanding concepts, analyzing the information, and creating new concepts [13].

2 Daphne Architecture

Daphne's main components and data flow are described below. Daphne has a web frontend that provides access to its capabilities and acts as the main User Interface (UI) for the system. A screenshot from this interface can be seen in Fig. 1. There are 3 main areas in the interface. The left area has a menu with all the available functions; the center area contains the design space plot—which allows for design space exploration—and the different tools available to the user, including a Design Builder; and the right area has the chat history between Daphne and the user.

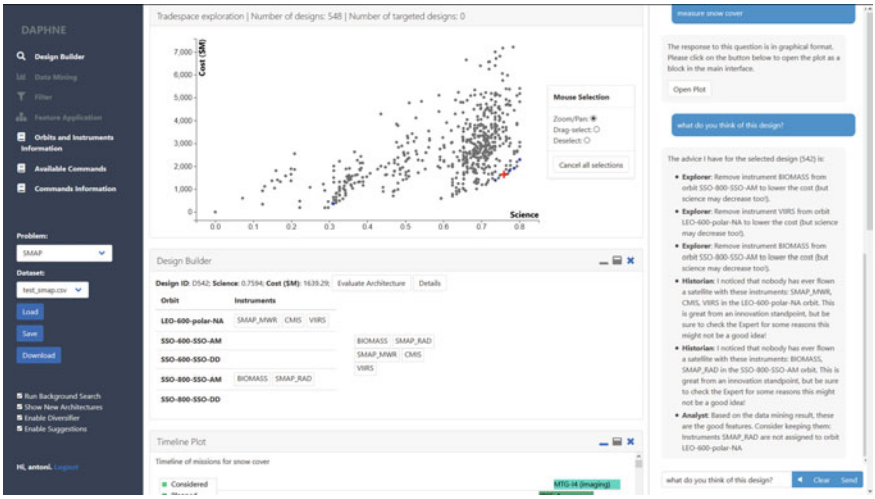

Fig. 1 Daphne's interface

The frontend is in charge of transmitting the user requests to the Daphne Brain, a web server in charge of forwarding all requests to the correct service. Requests can be natural language requests (text or voice) or classical interactions such as mouse clicks, hovering, or drag&drop.

Each request is processed by the Brain and sent to one of many *roles*, which are small programs that are in charge of handling groups of similar requests. All roles are also capable of being proactive and sending information to the user without a prompt, as described in [14].

There are 5 roles in Daphne. The first one is the Engineer, which is in charge of answering questions as a domain expert, as well as handling evaluation of the designs using models. Both functions of this role are supported by the VASSAR backend [15], a rule-based system for evaluating the performance and cost of Earth observation missions. The second one is the Analyst, whose job is to mine the dataset for knowledge on the shared features of designs in regions of interest in the objective space. This is supported by the iFEED backend [16], which searches for if—then rules that best explain a user-defined design region using a variety of rule mining algorithms. The third one is the Explorer, which controls a background search for better designs. The search is performed through the algorithm described in [17], which is an extension of a multi-objective genetic algorithm to include domain knowledge to make it more efficient. The fourth one is the Historian, which takes questions about past and current existing missions and answers them based on the data in the CEOS database.[1] The fifth and final one, the Critic, takes a design as

[1] http://database.eohandbook.com/.

an input and gives feedback on that design to the user. This feedback comes from all the other roles, but is synthetized in a few sentences from each role.

3 Experimental Design

To study the effects on learning and performance of using Daphne in the Earth observation task, we conducted a study at Texas A&M University with a diverse STEM student population. To address concerns about the measure for learning used in [10], we used a new, more holistic measure for learning in this study, which is described in the Dependent Variables subsection. We hoped to see that the more users interacted with Daphne, the more they learned. Also, while it was not the focus of the study, we explored if certain roles or functions of Daphne help users more than others in relation to both learning and performance.

Thus, we set the following hypotheses for the experiment:

- **H1:** There is a positive correlation between degree of Daphne usage and learning about the problem.
- **H2:** There is a positive correlation between degree of Daphne usage and performance on the design task.
- **H3:** There is a difference in the task performance when using Daphne as a Peer vs Daphne as an Assistant.
- **H4:** There is a difference in learning when using Daphne as a Peer vs Daphne as an Assistant.

4 Demographics

We recruited N = 26 Texas A&M Students from STEM degrees. Recruitment was through mass email on the university network, social media posts, and as part of a capstone design class for Aerospace Engineering students. All participants were promised a $15 gift card for a major online outlet as a token of appreciation for participation in the experiment. The main demographics are summarized below:

- Age range: 20–33 years old
- Gender: 20 identified as Male, 6 as Female, 0 as Others
- Current degree: 14 were BS students, 7 were MS students, 4 were PhD students, and 1 was a postdoctoral researcher
- Major: 12 were studying an Aerospace Engineering degree, while the rest were from various disciplines of STEM
- Prior Experience in Satellite Design: 7 subjects had previous experience, while 19 did not

Table 1 Features available in each condition

Capabilities	Condition 1—assistant	Condition 2—peer
Design space exploration	√	√
Design building	√	√
Explorer	√	√
Engineer	√	
Analyst	√	
Historian	√	
Critic		√

5 Experiment Protocol and Conditions

After signing all relevant IRB forms, each test subject sat down on a computer provided by us. Then, the subject was exposed to a tutorial explaining the protocol being described here and how to use all the functions of Daphne. The tutorial was interactive and thus had no time limit. This made the experiment have a variable duration, but we noticed in past experiments that limiting the tutorial time hurt performance. Once the tutorial was done, each test subject had to solve the design task under two different conditions, Peer vs Assistant. Each condition's available features in Daphne are detailed in Table 1 below. Participants were given 15 min to solve the task for each condition. After each task, the test subject was asked to complete a learning test, which was not time limited. At the end of the experiment, a semi-structured exist interview as conducted where subjects were asked to give their opinion on the experiment, the tool, and the task, to gather feedback for improving the system and the experiment.

Each subject performs two tasks (one per condition). The experimental design is between-subjects for H1 and H2 and within-subjects for H3 and H4. Each participant solved a problem of similar difficulty for each condition, and the order in which the conditions were given to the user was randomized to decrease the learning effect. All interactions with Daphne were recorded, from questions asked through the natural language interface to button clicks and hovering.

6 Task Details

The task given to the test subjects was the same as in [10] in order to allow for comparisons. Subjects were asked to design a satellite system to monitor soil moisture. They were given a set of 5 candidate orbits (e.g., different altitudes and inclinations) and a set of 5 candidate instruments (e.g., different types of infrared

and microwave sensors), and were asked to assign instruments to orbits with no constraints: every instrument can be in any subset of orbits, including none and all of them. The VASSAR backend [15] was used to assess the scientific value and cost of each design. Specifically, test subjects were tasked with finding a set of designs that push the boundary of the cost-science tradeoff (more formally the Pareto front) for costs between $800M and $4000M.

7 Dependent Variables

1. **Performance**: The true Pareto front for this design task is not known, so in order to measure performance we found an approximation of this optimal set by running a multi-objective genetic algorithm [17] for 10,000 evaluations. With this reference set, we defined the performance in the task as the distance between the user's Pareto front and the "reference" one found with the genetic algorithm. This distance was measured through the Hyper-Volume (HV) metric, a well-known metric in multi-objective optimization. We normalized the metric by bounding it between 0, if the subject's HV is the same as the starting one, and 1, if it is as good as the HV of the reference set.
2. **Learning**: One of the main limitations in past experiments was the metric for measuring learning. For this paper, we build on a study by Bang and Selva [13] on measures of learning for tradespace exploration problems. Their conclusion is that a learning test must target different cognitive processes such as remembering, understanding, analyzing, and creating. To do this, we defined three tests. The first one consists of 12 identification questions, where for a design chosen from the dataset the user was asked whether that design is close to the Pareto front or not. The second test also has 12 questions. For each question, the subject was asked to find the highest science design out of two designs that have a similar cost. For both tests, the test subjects were also asked to rate their confidence in their answers. Finally, we asked each subject three subjective questions on learning, to measure their perception of their own learning.
3. **Usability**: We conducted a standard usability survey after each task: The System Usability Scale (SUS) [18]. It consists of 10 Likert items, and has been validated in a large number of software usability studies, including intelligent systems.
4. **Trust**: We conducted a standard trust in automation survey after each task: the Jian's Trust in Automated Systems Scale [19]. It consists of 12 Likert items. This survey has been validated in a multitude of studies on automation, including VAs.

8 Results

In order to test both H1 and H2, we collected many usage statistics from each test subject during the 15 min they were performing the task: number of questions asked to Daphne, number of designs evaluated, and number of interactions. More detailed information was also recorded such as the number of interactions with each role (Critic, Engineer, Analyst, Historian), number of designs found by the Explorer vs the subject, etc. Then, the dependent variable data were separated in two groups based on usage (more usage versus less usage) and tested for difference in means. A selection of the results is plotted below. For the sake of brevity, Fig. 2 only details the interesting results for H1, while Fig. 3 only represents the interesting results for H2. Most variables had no correlation or trend and are not plotted. As a disclaimer, results from 8 test subjects were omitted because the Explorer did not work for those users and thus their scores could not be fairly compared to the others.

The plots in Fig. 2 show a trend of increased learning with increased usage, but the p-values for the t-tests are 0.14, 0.15, and 0.13 for #questions, #designs, and #interactions respectively, which are not significant.

The plots in Fig. 3 also show a trend of increased performance with increased usage, albeit weaker than that of the learning. The p-values for the t-test are 0.22, 0.42, and 0.76 respectively.

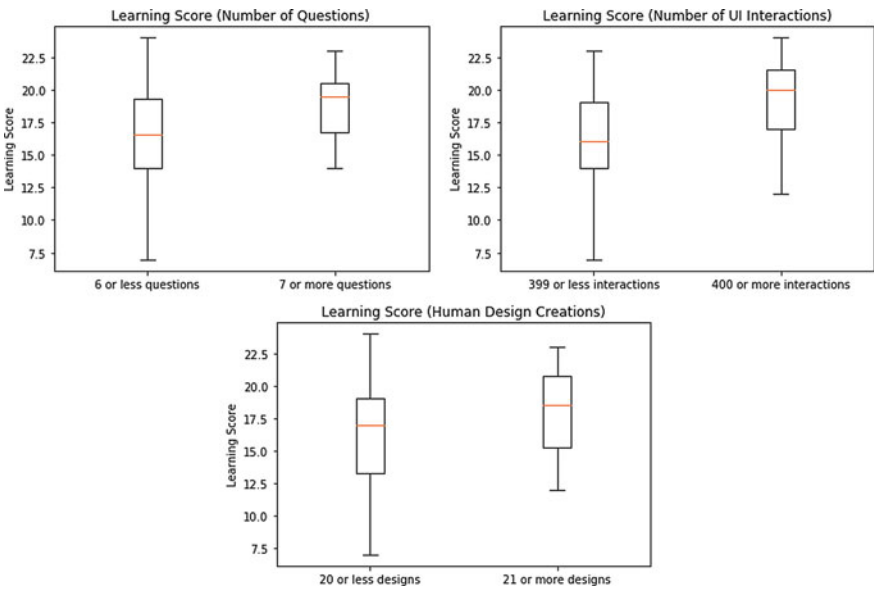

Fig. 2 Correlations between usage (#questions, #interactions, #designs evaluated) of the Daphne VA and learning

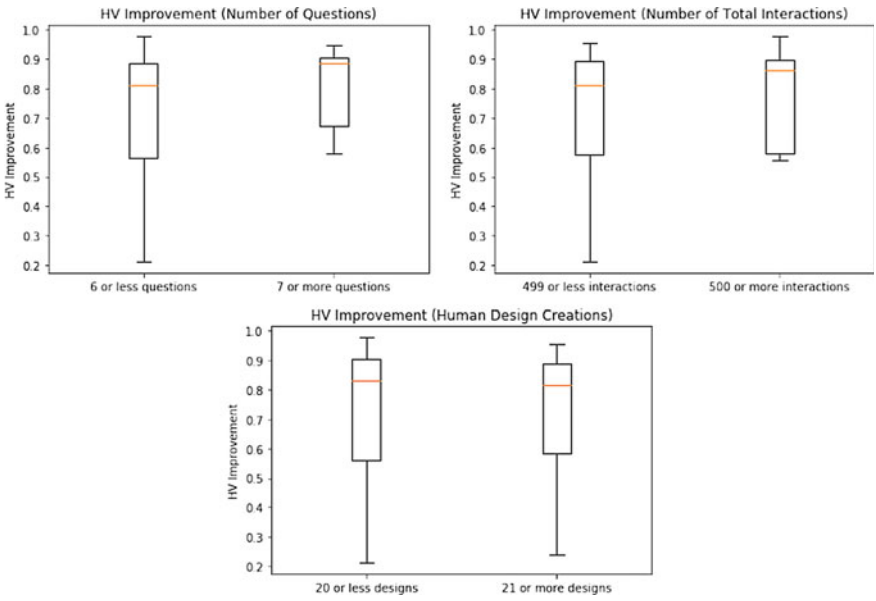

Fig. 3 Correlations between usage and performance

Finally, we also found an interesting relationship between the perceived usability (U), perceived trust (T), learning (L), and performance (P), which are detailed in Fig. 4 below. In this case, the p-values for the null hypothesis of zero slope (or no correlation) are 0.12 (L vs T), 0.00035 (L vs U), 0.07 (P vs T), and 0.57 (P vs U).

In order to test for H3 and H4, we compared the performance on the tasks and the learning scores of users when they used Daphne as a Peer vs Daphne as an Assistant. The distributions of results are shown in Fig. 5. They show no appreciable difference, and the p-values for the t-test confirm it, with values of 0.20 for the performance and 0.75 for learning.

9 Discussion

The results support H1 and H2, but not H3 and H4. The first two hypotheses are not supported with much strength, especially the second one. The trend that can be observed in most variables plotted for both H1 and H2 is that the more a user interacts with Daphne, the higher the lower bound is for both learning and performance. Some users are able to get great results with few interactions with the system, but having this lower bound raised by simply using the system more for the same amount of time is a result worth pointing out. If further studies can confirm this trend, we have an actionable way of fostering good learning and performance.

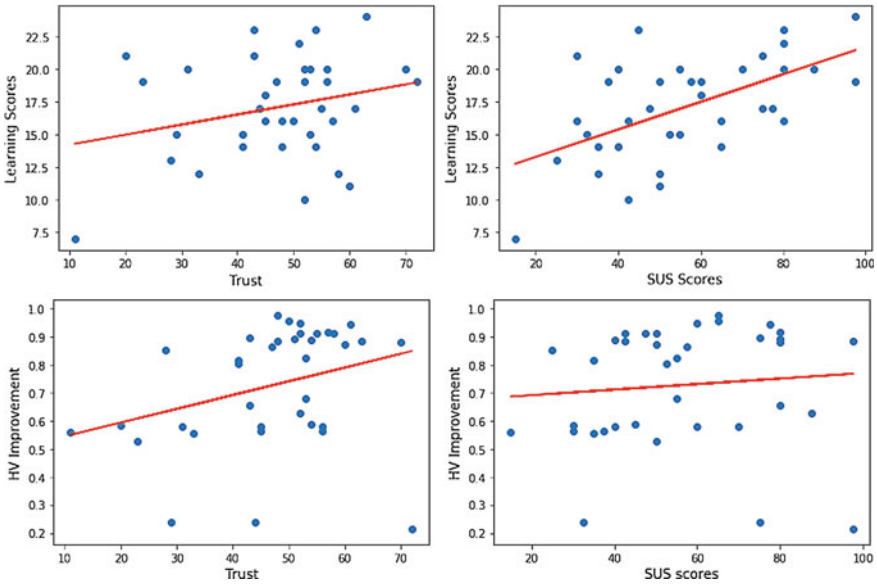

Fig. 4 Correlations between trust, usability, learning, and performance

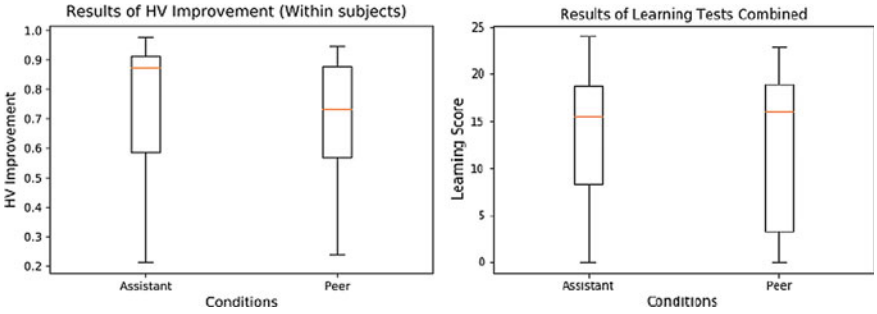

Fig. 5 Distributions of task performance and learning for each condition

This trend is also what one would expect: the more someone uses a system, the more proficient they become at it, so they also get more out of it.

We observe a strong correlation between the perceived usability of Daphne and how much learning there is according to the learning test score. It makes sense that the more a user learns, the more it finds Daphne usable, and vice versa.

Similarly, we observe that trust is correlated with performance. Although the relationship is not as strong as that of usability and learning, it seems that again, performing well with the system leads to higher trust scores with it. A further causality study could confirm these two findings.

We learn from the results that the tools available to the designer under each condition improve learning or performance by the same amount. A future study should look at whether the combination of the two roles (Peer + Assistant) results in higher learning or performance than either.

As far as time spent using different functionalities, users spent the most time using functions in the graphical UI, including both Design Space Exploration and Design Creation. These are followed in usage by backend roles such as the Analyst role and the Peer role. The Historian role was not used at all, and the Engineer role went almost unused by most users.

As the results lack significance, we also studied the qualitative feedback from the users' exit interviews. One consistent piece of feedback both in the interview and usage metrics is that subjects preferred roles such as the Data Mining and the Critic when compared to other roles such as the Engineer and Historian. Some subjects mentioned that these roles are more efficient to use in time-constrained situations such as the one in this experiment. The same effect was not seen when analyzing data from a previous experiment with subject matter experts; in fact, subjects familiar with satellite design mentioned that the Engineer tools were more helpful to them. A future experiment comparing the usage patterns of expert vs non-expert populations is needed to learn more about the trends we are seeing here, cater Daphne to the users that will end up using it, and learn whether it is appropriate to use students as subjects for further experiments.

10 Conclusion

This paper described an experiment to try to improve our understanding of the relation between key parameters in human-machine collaborative design space exploration. Specifically, we measure how using a VA may improve learning and performance in design space exploration. The main takeaways from this experiment are that increased interaction is linked to increased performance and learning, and that trust, usability, performance and learning tend to go hand in hand. We also found that STEM students (not real designers) prefer features that help them synthesize large amounts of data, as they spent more time using those features than others.

This study is not without limitations. Most results are not statistically significant, so more experiments are needed in order to confirm or deny the trends we have seen. Another important limitation is the allotted time for each experiment, which can be too short to both perform well on the task and learn meaningful facts about it. This means the findings and recommendations in this paper can be proven false in the future and should be tested independently. Further research is also warranted to understand the differences between non-experts (students) and expert practitioners in their usage of VAs and their various roles for design space exploration.

References

1. Myers K, Berry P, Blythe J, Conley K, Gervasio M (2007) An intelligent personal assistant for task and time management. AI Mag 28(2):47–62. https://doi.org/10.1609/aimag.v28i2.2039
2. Guo T, Lohan DJ, Cang R, Ren MY, Allison JT (2018) An indirect design representation for topology optimization using variational autoencoder and style transfer. In: 2018 AIAA/ASCE/AHS/ASC structures, structural dynamics, and materials conference, p 804
3. Hanna Landry, L, Cagan J (2011) Search strategies in evolutionary multi-agent systems: the effect of cooperation and reward on solution quality. J Mech Des 133(6)
4. Selva D, Cameron B, Crawley EF (2014) A rule-based method for scalable and traceable evaluation of system architectures. Res Eng Des 25(4):325–349. https://doi.org/10.1007/s00163-014-0180-x
5. Knerr N, Selva D (2016) Cityplot: visualization of high-dimensional design spaces with multiple criteria. J Mech Des 138(9):1–53. https://doi.org/10.1115/1.4033987
6. Van Horn D, Olewnik A, Lewis K (2012) Design analytics: capturing, understanding, and meeting customer needs using big data. In: ASME 2012 international design engineering technical conferences and computers and information in engineering conference, pp 863–875
7. Ferguson G, Allen JF (1998) TRIPS: an integrated intelligent problem-solving assistant. In: AAAI/IAAI, pp 567–572
8. Egan P, Cagan J (2016) Human and computational approaches for design problem-solving. In: Cash P, Stanković T, Štorga M (eds) Experimental Design Research, pp. 187–205. Springer, Cham. https://doi.org/10.1007/978-3-319-33781-4_11.
9. Hoffman G, Breazeal C (2004) Collaboration in human-robot teams. In: AIAA 1st intelligent systems technical conference, p 6434
10. Viros i Martin A, Selva D (2019) Daphne: a virtual assistant for designing earth observation distributed spacecraft missions. IEEE J Sel Top Appl Earth Obs Remote Sens
11. Viros Martin, A, Selva, D (2020) Explanation approaches for the Daphne virtual assistant. In: AIAA Scitech 2020 Forum, p 2254
12. Krathwohl DR, Anderson LW (2009) A taxonomy for learning, teaching, and assessing: a revision of bloom's taxonomy of educational objectives. Longman
13. Bang H, Selva D (2020) Measuring learning to assess effectiveness of knowledge discovery tools in design space exploration. In: Accepted at IDETC/CIE 2020 (2020)
14. Virós, A, Selva D (2019) From design assistants to design peers: turning Daphne into an AI companion for mission designers. In: AIAA Scitech 2019 Forum. https://doi.org/10.2514/6.2019-0402
15. Selva D (2014) Knowledge-intensive global optimization of Earth observing system architectures: a climate-centric case study. SPIE Remote Sens 9241:1–22. https://doi.org/10.1117/12.2067558
16. Bang H, Selva D (2020) Discovering generalized design knowledge using a multi-objective evolutionary algorithm with generalization operators. Expert Syst Appl 143: 113025
17. Hitomi N, Bang H, Selva D (2018) Adaptive knowledge-driven optimization for architecting a distributed satellite system. J Aerosp Inf Syst 15(8):485–500
18. Brooke J (1996) SUS-A quick and dirty usability scale. Usability Eval Ind 189(194):4–7
19. Jian J-Y, Bisantz AM, Drury CG (2000) Foundations for an empirically determined scale of trust in automated systems. Int J Cogn Ergon 4(1):53–71