Empirical Evaluation of State Event Fault Tree and Dynamic Fault Tree for the Safety Analysis of Ambient Assisted Living

Adrien Mouaffo, Kavyashree Jamboti, Davide Taibi
Software Engineering Research Group
University of Kaiserslautern
Kaiserslautern, Germany
{adrien.mouaffo | jamboti | taibi}@cs.uni-kl.de

TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

AG·SE

Abstract

Most of the evolution in ambient assisted living is due to embedded systems that dynamically adapt themself to react to environmental changes or component/subsystem failures to maintain a certain level of safety. Following this evolution fault tree analysis techniques have been extended with concept for dynamic adaptation but resulting techniques such as dynamic fault trees or state event fault trees analysis are not widely used as expected.

In this report we describe a controlled experiment to analyze these two techniques with regard to their applicability and efficiency in modeling dynamic behavior of ambient assisted living systems.

Results of the experiment show that DFTs are easier and more effective to use, althought they produce better results (models) with SEFTs.

Table of Contents

List of Figures

List of Tables

1 Introduction

Embedded system usage is growing with fields like ubiquitous computing. One application domain for ubiquitous computing is ambient assisted living. Ambient assisted living systems make usage of dynamic adaption to environmental change or component/subsystem failures for remaining safe. Following this evolution, fault tree analysis techniques have been extended with concept for dynamic adaptation but resulting techniques such as dynamic fault tree or state event fault tree analysis are not widely used as expected.

In this report we describe a controlled experiment to analyze these two techniques with regard to their applicability and efficiency in modeling dynamic behavior of ambient assisted living systems.

The remainder of this report is structured as follows: Section 2 provides background knowledge on AAL and safety analysis, In section 3 design and execution of the experiment are explained, then section 4 describes performed analysis, we discussed analysis results in section 5 and show how we avoid validity threats in section6, finally we conclude in section 7.

2 Background

In this section we give an overall background which to understand the performed study. We first describe what ambient assisted living is and why ambient assisted living systems are important. Then we describe safety analysis.

2.1 Ambient Assisted Living

Due to the the progress in medical treatment and pharmacies in most industrialized countries life expectancy has dramatically increased in Europe and the western hemisphere. This growing share of elderly people with assistance needs, lead to dramatic effects on public and private health care, emergency medical services, and the quality of life of individuals themselves. An exemplary study in the district of Kaiserslautern, Germany, shows that 44% of Emergency Medical Services' (EMS) system resources are dedicated to patients older than 70 years of age [1]. Assisted living helps elderly people to cope with their daily chores and challenges when becoming older. Kleinberger et al. listed three important characteristics of assisted living systems [1]:

- They have to be ambient and unobtrusive to reach a high acceptance.
- They have to adapt themselves to changing personal situations or capabilities of the individual and the environment to fulfill individual needs.
- They have to provide their services in an accessible way to enhance usability.

To achieve these goals Ambient Intelligent (AmI) is a promising approach. Systems that render their service in a sensitive and responsive way and are unobtrusively integrated into our daily environment are referred to as being ambient intelligent [1].

According to Oppermann et al. [2], AmI systems represent a new generation of systems that show the following characteristics:

- **Invisible,** i.e., embedded in clothes, watches, glasses, etc.,
- **Mobile,** i.e., being carried around,
- **Spontaneous (ad hoc) communication** among the nodes,
- **Heterogeneous and hierarchical,** i.e., they comprise different kinds of system nodes regarding their computational power and rendered functionality,

- **Context-aware,** i.e., they are aware of their local environment and spontaneously exchange information with similar nodes in their neighbourhood,
- **Anticipatory,** i.e., acting on their own behalf without explicit extrinsic requests,
- **Natural communication** with users by voice and gestures instead of keyboard, mouse, or text on screens,
- **Natural interaction** with users by means of devices they are used to, e.g., clothing, watches, TV, telephone, household appliances. To this end the devices will be equipped with some kind of intelligence.
- **Adaptive,** i.e., capable of reacting to all abnormal and exceptional situations in a flexible way.

AmI-based assisted living systems called Ambient Assisted Living (AAL) systems consist of various sensors, actuators and software components integrated into everyday items or worn/used by patients. They can be classified in three categories (see Figure 1) [3, 1]:

- **Emergency treatment:** considered as the kernel of any living assistance system, it aims at the early prediction of and recovery from critical conditions that might result in an emergency situation and the safe detection and alert propagation of emergency situations.
- **Autonomy Enhancement:** Autonomy enhancement services are services that make it possible to abandon previous manual care given by medical and social care personnel or relatives, and replace it by appropriate system support.
- **Comfort:** Comfort services cover all areas that do not fall into the first two categories above. It is clear from the discussion that comfort services do not have the same importance and social impact as the other two categories. But they might increase the acceptance of AAL systems on the mass market, especially for customers that currently do not depend on the autonomy and emergency treatment, but will probably do in some years.

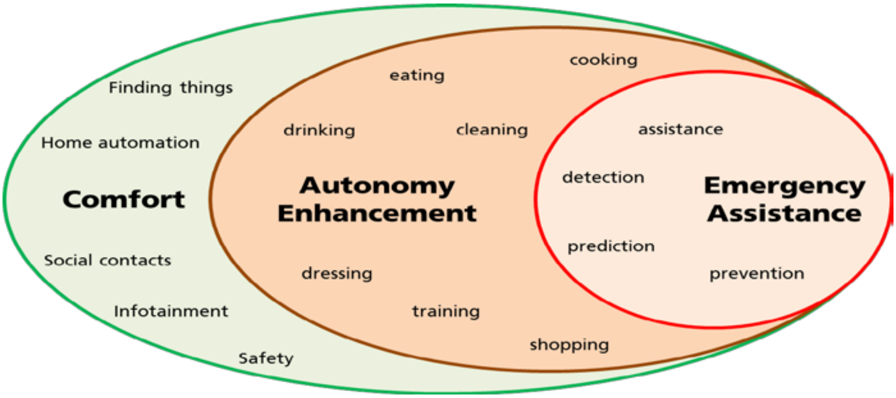

Figure 1 AAL classification

Like all AmI systems, AAL systems are invisible, mobile, context-sensitive, proactive, adaptive, and of course in communication with the users, above all, they must be safe with regard to the people using the system, ie the user must not suffer from injuries in case of malicious attacks to the - or faults of the system. One challenge in the development relates to compliance with certain safety requirements of such open and distributed systems. In the research area of Ambient Assisted Living (AAL), the University of Kaiserslautern and the Fraunhofer IESE have developed systems, which allow through information and communication technology, older people or people with disabilities to live independently in their own homes instead of living in nursing homes (Alter-Wohnheim"). The goal is to develop intelligent network systems, which get information from a number of discreet sensors installed in the home and are able to analyze the situation of the assisted person in the apartment and to act accordingly. A central role is played by the support of emergency situations.

2.2 Safety Analysis

In this section we introduce some basics on safety of a system. We first introduce some basic vocabulary namely the definition of faults, failures, hazards and accidents [4]. Afterwards we describe safety analysis techniques based on fault tree models.

2.2.1 Faults, Failures, Hazards and accidents

The terms failure and fault are the key to any understanding of FT construction. Yet they are often misused. One of them describes the situation(s) to be avoided, while the other describes the problem(s) to be circumvented. In this section we briefly recall the main definitions for performing safety analysis using FT and CFT. For further information we refer to [5, 6, 7].

Failure

Each behavior of a system that differs from the ambient conditions specified behavior although the environmental conditions are specified correctly is called a failure.

Fault/Error

A fault is a static event in a system that may cause a failure. There are many different definitions for a fault. Some of them even differentiate between error and fault. But the main consensus is found in the difference between fault and failure. A fault is a deviation from the specification such as incorrect design or incorrect usage. A failure is a state of a system. A fault may but needn't cause a failure. If a failure is present then it must be cause by one or more faults. Also a fault may be caused by other faults.

Accident, Mishap

An accident is an undesired event that destroys or affects goods such as life or health of humans, economical goods, or the environment.

Hazard

A hazard is a state of the system in scope and its environment in which the occurrence of an accident only depends on uncontrollable influences. A hazard is for example, an open gate in the presence of an arriving train. Whether an accident occurs depends only on whether the driver of a car near the gate is alerted or not.

2.2.2 Fault Tree Analysis

Fault trees [8, 9, 10] are constructed using a backward searching technique starting with a top event. The causes identified are combined using boolean gates. After its construction, it can provide quantitative results such as the top event probability or qualitative results in the form of Minmal Cut Sets (MCS). A MCS signifies a set of events where the nonoccurrence of even one event prevents the top event from occurring. The MCS can be ranked according to the number of events comprising them and the ones with less number of events need to be ensured that their occurrence probabilities are reduced or eliminated. In cases where the MCS consists of just one event called a single point failure, special attention must be given in order to ensure that it does not occur or its chances are minimized.

2.2.3 Component Fault Trees

In the previous section we recalled conventional FTs. The modeling of CFTs is a modularization technique to handle FTs for huge systems that consist of more than one component. The components are connected in a functional network via signal ports and the top events of the CFTs correspond to failure modes of the output signals. CFTs are similar to classical FT with some differences. They may contain more than one top event and one basic event may also be connected to more than one logical gate. CFTs have been developed in 2003 by P. Liggesmeyer, O. Mäckel and B. Kaiser, see [6] for further details.

2.2.4 Dynamic Fault Trees

As the name suggests, this techniques [11] enable one to analyze top events for dynamic systems where the notion of spare components is predominant. DFTs enable modeling of stochastically dependent events (failures of spare parts and triggered events) and sequencing by using new types of gates listed below:

- PAND(Priority-And)
- SEQ(Sequence-Enforcing)
- FDEP(Functional Dependency)
- CSP(Cold spare), WSP(Warm spare) or HSP(Hot Spare)

It is important to note that DFTs are analyzed by analyzing underlying Markov chains that capture the sequencing and stochastic dependence of events.

2.2.5 State/Event Fault Trees

State/Event Fault Trees (SEFTs) [12] build on CFTs [6] which are an elegant approach to build fault trees based on failures of the components of a system. A CFT overcomes the drawbacks of the traditional fault tree which only conveys how a failure can occur, but does not specify which components influence each other in a manner that the failure occurs. CFTs can be easily reused as they have clear decomposition semantics based on system architecture. Though CFTs overcome some of the drawbacks of traditional fault trees, they are incapable of handling some other issues of fault trees such as sequence and timing issues of fault tree events. CFTs cannot handle stochastic dependence and cannot be integrated with state-based design models showing the behavior of the system. SEFTs have been designed to overcome the above problems. They allow the modeling of failure of a component showing the internal safety relevant state changes. Unlike traditional FTs or CFTs they make a clear distinction between a state and an event. In the context of SEFTs, a state is defined as the collectivity of the variable properties of a component that are relevant to its behavior and its reaction to external events and an event is defined as a sudden phenomenon without temporal expansion in the context of discrete event systems. A state or event occurrence in one component can trigger state changes in another component. SEFTs enable the use of a wide range of gates which need not be just boolean operators provides by traditional FTs, gates in SEFTs can be made of boolean operators and state-based models which allow modeling of the order and timing of the occurrence of states and events in an SEFT. Some of the gates used in an SEFT are:

- AND(with n state inputs)
- AND(with n state inputs and one event input)
- OR(with n state inputs)
- OR(with n event inputs)
- History-AND
- Priority-AND

SEFTs are quantitatively analyzed by translation to Petri Nets. The top event probability can be calculated by calculating the probability for the corresponding place in the Petri net.

3 Experiment Design and Execution

In this section we specify the goal of the experiment, describe the design used for the experiment and the procedure followed for its execution.

3.1 Goal

We specify in this section the goal of the controlled experiment. The high level goal is firstly specified using the GQM goal specification template. After that we derive the questions and metrics related to the goal and describe them using a tree structure.

3.1.1 GQM Goal specification

Following the GQM goal specification construct, the goal of the experiment is to:

Analyze state/Event Fault Tree (SEFT) and dynamic Fault Tree (DFT) for the purpose of understanding and comparing their applicability and efficiency with respect to the modeling of safety related aspects of Ambient Assisted Living systems.

Figure 2 presents the corresponding GQM Tree. On the tree metrics used for answering the questions have been defined based on those provided by the technology acceptance model [13, 14]:

- **Completeness:** measures the capability of a method to completely model all aspects of the system.
- **Easiness:** measures the effort needed for building the model.
- **Understandability:** measures the effort needed to understand the models built with the technique in relation to the failure logic.
- **Time needed:** measures the time needed for building the models
- **Quality of produced models:** measures how good produced models are.
- **Effort expectancy:** measures the degree of ease associated with the use of the technique.
- **Attitude toward using the methodology:** measures the overall affective reaction to using the technique.
- **Self efficacy:** measures the degree to which the subject believes that they will better perform if they have some help.
- **Performance expectancy:** measures the degree to which the subject believes that using the technique will help him to attain gains in job performance. This metric was only valid for people

who are professionnaly active (researchers or assistant researchers).

Measurements are applied for each group. And for each metric a test is perform to compare results between both groups.

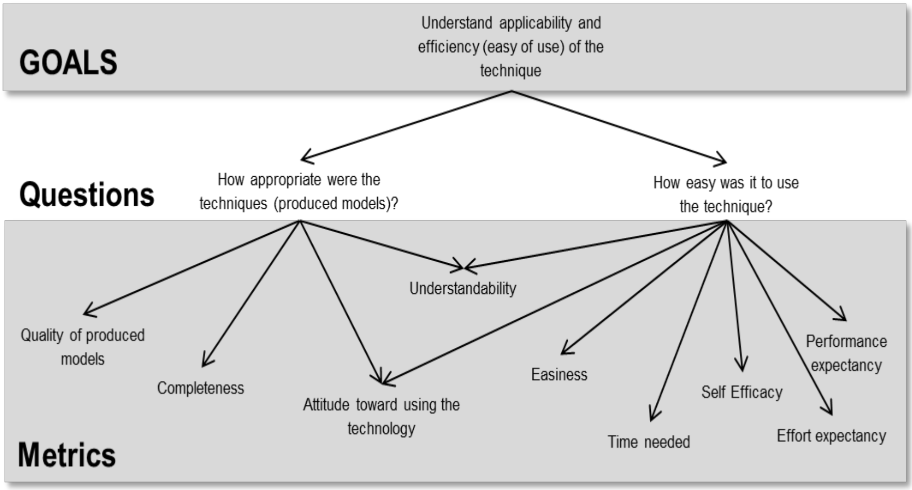
GOALS
Understand applicability and efficiency (easy of use) of the technique
Questions
How appropriate were the techniques (produced models)?
How easy was it to use the technique?
Metrics
Understandability
Quality of produced models
Completeness
Attitude toward using the technology
Easiness
Time needed
Self Efficacy
Performance expectancy
Effort expectancy

Figure 2 GQM Tree Specification

3.1.2 Hypotheses

In order to compare assement results for both groups we specified following hypotheses:

- **Null hypothesis:**
  H0: The score for both groups are similar.
- **Alternative hypotheses**
  H1: The score for both groups are differents and the score for the SEFT group is better than the score for the DFT group.

3.2 Participants

Participants are 8 students from the lecture "Empirical model building" taught at the TU Kaiserslautern and 6 researchers from the research group Software Engineering: Dependability of the TU Kaiserslautern. The selection criterion for students was to have safety analysis knowledge and also be motivated to be part of a real experiment out of the theoretical boundary of the lecture. Researchers were interested in the experiment topic. Subjects were randomly assigned to 2 groups of 7 subjects each. Each group containing 3 Researchers and 4 students.

A questionnaire was used for knowing the background of each participant. Table 1 and Table 2 show the repartition.

DFT	SEFT
Software Engineering (2) Visualization – safety analysis (1) HCI and Software Engineering (1) Computer Science/Engineering (2)	Computer Science/Engineering (3) Software Engineering (3)

Table 1 Majors

DFT	SEFT
Diplom Informatiker (3) Mechanical Engineering with applied computer science (Diploma) (1) B. Sc. Computer Science (1) M. Sc. Computer Science (1) Computer Science	Student Technoinformatik (1) M. Sc. Software Engineering (1) B. Sc. Computer Science (3) M. Sc. Information retrieval (1)

Table 2 Academic grade

3.3 Material and instruments

Participants of each group were trained for their respective techniques before the experiment. We use a questionnaire for getting their feedback on the training. After the training participants receive material describing the system used during the experiment. Each task was described and at the end of the task description a questionnaire was added for getting the impression of each participant after the executuion of the given task. At the end of the experiment each participant then has to fill in a debriefing questionnaire.

3.4 Execution

Here we report details on the design and precedures.

3.4.1 Design and procedure

The subjects were randomly distributed in two separated groups, one for subjects applying SEFT and the other one for subjects applying DFT (Table 3).

	SEFT	DFT
AAL System	Group 01	Group 02

Table 3 Experiment design

3.4.2 Procedure

Table 4 shows the chronological procedure of the experiment. Because we only had one trainer and we assigned a different room for each group, we needed to start the experiment at different time. T

Order	Room SEFT	Room DFT
1	Pre-questionnaire	
2	Training: Introduction of SEFT concepts	Pre-questionnaire
3	Feedback session on training	Training: Introduction of DFT concepts
4	Introduction to the study	Feedback session on training
5	Desccription of the AAL system example	Introduction to the study
6	Distribution of material	Desccription of the AAL system example
7	Execution of tasks 01	Distribution of material
8	Execution of tasks 02	Execution of tasks 01
9	Execution of tasks 03	Execution of tasks 02
10	Debriefing	Execution of tasks 03
11	-	Debriefing

Table 4 Experiment procedure

3.4.3 Deviation from plan

Following deviations were observed during the experiment:

- Students were not as motivated as expected
- Questionnaire were not answered properly
- No all time information were provided
- Training tutorial took too long
- Too many paper material

4 Analysis

During and after the experiment we collected data to be analyzed for testing our hypotheses. In this section we report and analyze those data.

4.1 Data collection and aggregation

Subjects of each group have to perform 3 different tasks of growing complexity: Task 01, Task 02 and Task 03. After each task they answer a questionnaire of 11 questions. We will reference question Y of task X as TX.Y. E.g. T02.08 represents the 8th question of task 02.

After performing each task subjects have to answer a debriefing questionnaire with 16 questions on the technique they were applying. We will reference question Y of the debriefing questionnaire as D.Y. E.g. D.06 represents the 6th question of the debriefing questionnaire.

Table 5 shows how these questions are related to the respective metrics. Completeness, easiness, understandability, time needed and quality of produced models are calculated for each task.

Metrics	Questions
Completeness	TX.01: **I am sure that I was able to transfer the description from the system model completely to the DFT / SEFT.**
	TX.03: **I was able to identify appropriate gates for describing all failure logics. / I was able to identify the locations in the SEFTs that needed to be involved for doing the modifications.**
	TX.04: **I am sure that I was able to identify all changes that have to be made from the original DFT. / I am sure that I was able to identify all the involved locations in the SEFT.**
Easiness	TX.02: **It was easy for me to transfer descriptions of the system model to the DFT / SEFT.**
	TX.08: **The DFTs / SEFTs supported me during the accomplishment of the tasks.**
	TX.09: **It was easy for me to implement the modifications.**
	TX.10: **I was able to make the modifications with minor effort.**
	TX.11: **I was able to re-use a lot from the existing model during the modifications.**

Understandability	TX.05: **Because of the graphical representation of DFTs / SEFTs it was easy for me to keep the overview of the failure logic.**
	TX.06: **The relationship between the DFTs / SEFTs and system is easy for me to comprehend.**
	TX.07: **The DFT / SEFT methodology helped me to keep the overview of the failure logic.**
Time needed	Time needed for Task 1
	Time needed for Task 2
	Time needed for Task 3
Quality of produced models	Quality of produced model for Task 1
	Quality of produced model for Task 2
	Quality of produced model for Task 3
Effort Expectancy	D.01: **The DFTs / SEFTs methodology is clear and understandable.**
	D.02: **It was easy for me to work with the DFTs / SEFTs.**
	D.03: **I find the DFTs / SEFTs easy to use.**
	D.04: **Learning to use the DFTs / SEFTs was easy for me.**
Attitude toward using the method	D.05: **Using the DFTs / SEFTs is a good idea.**
	D.06: **The DFTs / SEFTs make work more interesting.**
	D.07: **Using the DFTs / SEFTs is fun.**
	D.08: **I like using the DFTs / SEFTs.**
Self- Efficacy	*I could complete a job or task using the Dynamic Fault Trees / State Event Fault Trees…*
	D.09: **… if there was no one around to tell me what to do as I go.**
	D.10: **… if I could call someone for help if I got stuck.**
	D.11: **… if I had a lot of time to complete the job for which the DFT was provided.**
	D.12: **… if I had just the built-in help facility for assistance.**
Performance Expectancy (answered only by researchers)	D.13: **I would find the DFTs useful in my work.**
	D.14: **Using the DFTs enables me to accomplish tasks more quickly.**
	D.15: **Using the DFTs increases my productivity.**
	D.16: **If I use the DFTs, I will increase my chances of getting a raise. (e.g., by being faster)**

Table 5: Relation between Metrics and questions

4.2 Analysis procedures

The answer for each question was given on a scale of 1 to 5:

- **1:** The subject strongly disagrees
- **2:** The subject disagrees
- **3:** The subject neither disagrees nor agrees
- **4:** The subject agrees
- **5:** The subject strongly agrees

For aggregating the answers of questions into metrics, we calculate the metric score as followed:

$$\text{Metric score} = \left(\sum_{i=1}^{n} \text{question}_i\right) / (n * 5)$$

Equation 1 Metric score calculation

The metric score is calculated per metric for each subject. It is a percentage value which expresses how close the score is to the ideal answer that is the subject strongly agrees about all questions which are related to the metric. Values for the metric score range from 0.2 if the subject strongly agrees for all questions and 1.0 if the subject strongly disagrees for all questions. If a score is less than or equal to 0.6 then the result is considered as negative.

For each metric, a descriptive statistic analysis is performed for giving quantitative statistical information abouth the metric. Then a hypothesis test is performed to gain confidence on the results (median) from the descriptive statistics.

For comparing the metric scores between both groups we first test the data for normality with a Shapiro-Wilk’s test. When scores for both groups are normally distributed we perform an Independent T-Test [t-test] for comparing the means. Else we perform a median test for comparing the medians.

4.3 Analysis results

We report in this section analysis results for each metric of the GQM tree (Figure 2).

4.3.1 Completeness

Completeness measures the capability of a method to completely model all aspects of the system.Completeness score for task X is calculated from the answer of TX.01, TX.03 and TX04.

Completeness scores decrease with complexity in both groups. For the 1st task the score is very high for both groups (Median = 0,933). In the 2nd task the score for the DFT group stays at a good level (Median = 0,86) when the score for the SEFT group decreases to Median=0,66. The situation observed in the first task for the SEFT group reproduces itself for the 3rd task in both groups, where the score for the SEFT group falls under 0,6.

For the first task completeness scores were similar for both groups, but for the 2nd and 3rd task subjects believe they achieve a better completeness in the DFT group. Statistical significance for our comparison could not be achieved for all quantitative computations.

Following we present in detail the analysis of completeness for each task.

Completeness for Task 01

As shown in Table 6 the completeness score for task 01 in both groups has a median of 0,9333. A detailed view of T01.01, T01.03 and T01.04 shows us that for all this questions the median is also similar. A hypothesis test confirms the results of median calculation for both groups with an acceptable confidence level (0,026 for DFT-group and 0,033 for SEFT group)

Descriptives				
	Group		Statistic	Std. Error
Completeness_Task1	DFT	Mean	,8952	,05607
		Median	,9333	
		Variance	,022	
		Std. Deviation	,14836	
		Minimum	,60	
		Maximum	1,00	
	SEFT	Mean	,8667	,08230
		Median	,9333	
		Variance	,047	
		Std. Deviation	,21773	
		Minimum	,40	
		Maximum	1,00	

Table 6: Completeness for task 01

Completeness scores in task 01 were not normaly distributed for both groups as assessed by the Shapiro-Wilk's test (Sig.= 0,021 for DFT group and Sig. = 0,003 for SEFT group). Therefore we perform a median test to compare the median obtained previously. The median test retain the null hypothesis: the median of completeness score in both groups are the same (Figure 3).

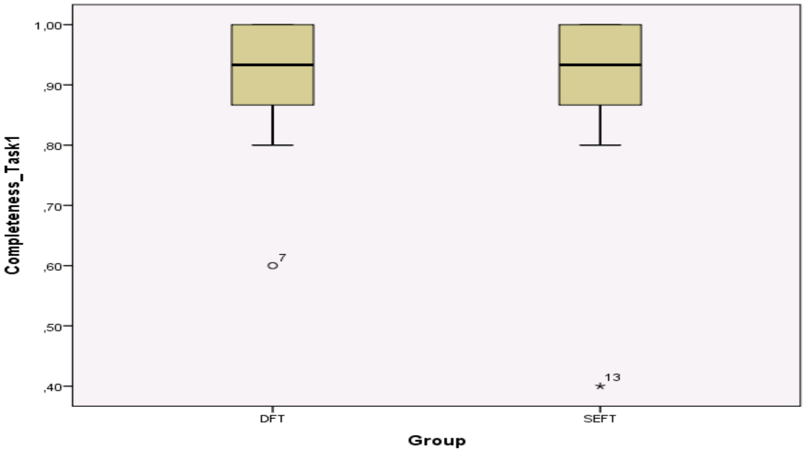

Figure 3: Boxplot completness in task 01

Completness for Task 02

As shown in Table 7 the completeness score for task 02 has a median of 0,8667 in DFT-Group and 0,6667 in SEFT group. A detailed view of T02.01, T02.03 and T02.04 shows us that the difference is on answers for T02.01 and T02.04 where the results a better for the DFT-Group. A hypothesis test confirms the results of median calculation only for the DFT-Group with an acceptable confidence level (0,027)

Descriptives				
	Group		Statistic	Std. Error
Completeness_Task2	DFT	Mean	,8571	,05331
		Median	,8667	
		Variance	,020	
		Std. Deviation	,14105	
		Minimum	,60	
		Maximum	1,00	
	SEFT	Mean	,7143	,07930
		Median	,6667	
		Variance	,044	
		Std. Deviation	,20981	
		Minimum	,40	

		Maximum	1,00	

Table 7: Completenees for Task 03

Completeness scores in task 02 were normaly distributed for both groups as assessed by the Shapiro-Wilk's test (Sig.= 0,362 for DFT group and Sig. = 0,804 for SEFT group). There were no outliers in the data, as assessed by inspection of the boxplot (Figure 4). Therefore we perform an independent-samples t-test to compare the mean obtained for both groups. Completeness score in task 02 for DFT-Group (Mean = 0,8571 ± 0,14) was 0,142 higher than completeness score in task 02 for SEFT-Group (Mean = 0,7143 ± 0,20). The mean difference (0,145) was not satistically significant (p=0,161).

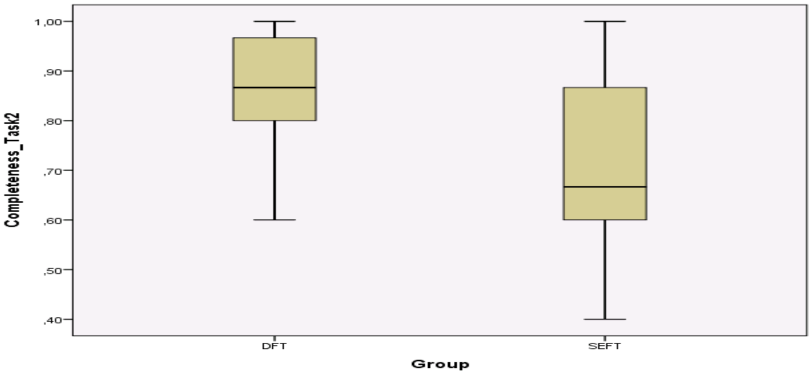

Figure 4: Boxplot completeness in Task 02

Completeness for Task 03

As shown in Table 8 the completeness score for task 03 has a median of 0,7333 in DFT-Group and 0,6667 in SEFT group. A detailed view of T03.01, T03.03 and T03.04 shows us that the difference is on answers for T03.01 and T03.04 where the results a better for the DFT-Group, although the answers for T03.03 are better for the SEFT-Group. A hypothesis test does not confirm the results of median calculation with an acceptable confidence level

Descriptives				
	Group		Statistic	Std. Error
Completeness_Task3	DFT	Mean	,6762	,08781
		Median	,7333	
		Variance	,054	
		Std. Deviation	,23231	
		Minimum	,40	
		Maximum	1,00	
	SEFT	Mean	,5619	,09299

		Median	,6667	
		Variance	,061	
		Std. Deviation	,24603	
		Minimum	,20	
		Maximum	,80	

Table 8 Descriptive statistics: Completeness for task 03

Completeness scores in task 03 were normaly distributed for both groups as assessed by the Shapiro-Wilk's test (Sig.= 0,379 for DFT group and Sig. = 0,188 for SEFT group). There were no outliers in the data, as assessed by inspection of the boxplot (Figure 5). Therefore we perform an independent-samples t-test to compare the mean obtained for both groups. Completeness score in task 03 for DFT-Group (Mean = 0,6762 ± 0,23) was 0,114 higher than completeness score in task 03 for SEFT-Group (Mean = 0,5619 ± 0,24). The mean difference (0,114) was not satistically significant (p=0,389).

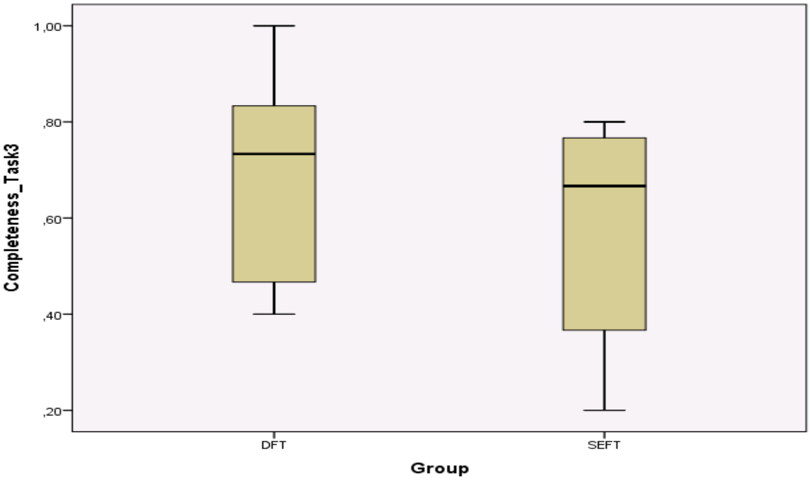

Figure 5 Boxplot for Completeness in task 03

4.3.2 Easiness

Easiness measures the effort needed for building the model. Easiness score for task X is calculated from the answer of TX.02, TX.08, TX.09, TX.10 and TX11.

Easiness scores decrease with complexity in both groups and is similar for the SEFT group in the 2nd and 3rd task. For the 1st task the score is very high for both groups (Median = 0,9). In the 2nd task the score for the DFT group stays at a good level (Median = 0,88) when the score for the SEFT group decreases to Median=0,68. The situation observed in the first task for the SEFT group reproduces itself for the 3rd task in the DFT groups (Median=0,6). We have to notice that the score in the 3rd task for the SEFT group is better than the score of the DFT group.

For the first task Easiness scores were similar for both groups, but for the 2nd task subjects believe they have a better Easiness in the DFT group. In the 3rd task subjects believe that they have a better Easiness in the SEFT group.

Statistical significance for our comparison could not be achieved for all quantitative computations.

Following we present in detail the analysis of Easiness for each task.

Easiness for task 01

As shown in Table 9 the easiness score for task 01 has a median of 0,96 in DFT-Group and 0,92 in SEFT group. A detailed view of T01.02, T01.08, T01.09, T01.10 and T01.11 shows us that the difference is on answers for T01.08 where the results a better for the DFT-Group. A hypothesis test confirms the results of median calculation with an acceptable confidence level (0,026 for DFT-Group and 0,027 for SEFT-Group).

Descriptives[a]

	Group		Statistic	Std. Error
Easiness_Task1	DFT	Mean	,9400	,03225
		Median	,9600	
		Variance	,006	
		Std. Deviation	,07899	
		Minimum	,80	
		Maximum	1,00	
	SEFT	Mean	,8686	,06801
		Median	,9200	
		Variance	,032	
		Std. Deviation	,17995	
		Minimum	,48	
		Maximum	1,00	

Table 9 Descriptive statistics for Easiness_Task 1

Easiness scores in task 01 were not normaly distributed for both groups as assessed by the Shapiro-Wilk's test (Sig.= 0,060 for DFT group and Sig. = 0,005 for SEFT group). Therefore we perform a median test to compare the median obtained previously. The median test retains the null hypothesis: the median of completeness score in both groups are the same (Figure 6).

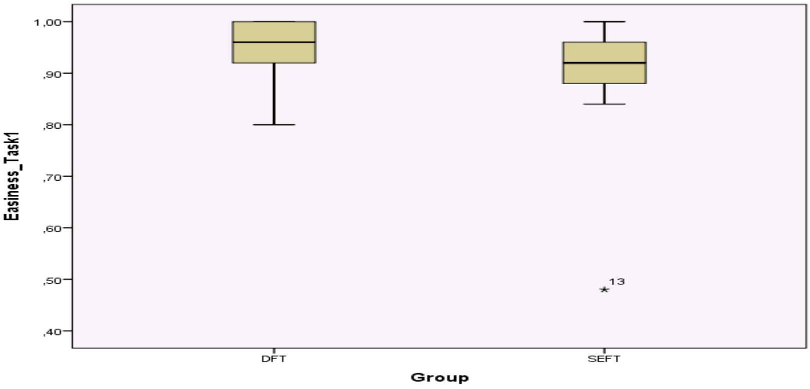

Figure 6 Boxplot for Easiness of Task 01

Easiness for Task 02

As shown in Table 10 the easiness score for task 02 has a median of 0,88 in DFT-Group and 0,68 in SEFT group. A detailed view of T02.02, T02.08, T02.09, T02.10 and T02.11 shows us that the difference is on answers for T02.02, T02.09, T02.10 and T02.11 where the results is better for the DFT-Group. A hypothesis test confirms the results of median calculation with an acceptable confidence level only for the DFT-Group (0,027).

Descriptives

	Group		Statistic	Std. Error
Easiness_Task2	DFT	Mean	,8400	,06294
		Median	,8800	
		Variance	,028	
		Std. Deviation	,16653	
		Minimum	,60	
		Maximum	1,00	
	SEFT	Mean	,7029	,05911
		Median	,6800	
		Variance	,024	
		Std. Deviation	,15639	
		Minimum	,52	
		Maximum	1,00	

Table 10 Descriptive Statistics: Easiness for Task 02

Easiness scores in task 02 were normaly distributed for both groups as assessed by the Shapiro-Wilk's test (Sig.= 0,193 for DFT group and Sig. = 0,393 for SEFT group). There were no outliers in the data, as assessed by inspection of the boxplot (Figure 7). Therefore we perform an independent-samples t-test to compare the mean obtained for both groups. Easiness score in task 02 for DFT-Group (Mean = 0,84 ± 0,166) was 0,137 higher than easiness score in task 02 for SEFT-Group (Mean = 0,70 ± 0,156). The mean difference (0,137) was not satistically significant (p=0,138).

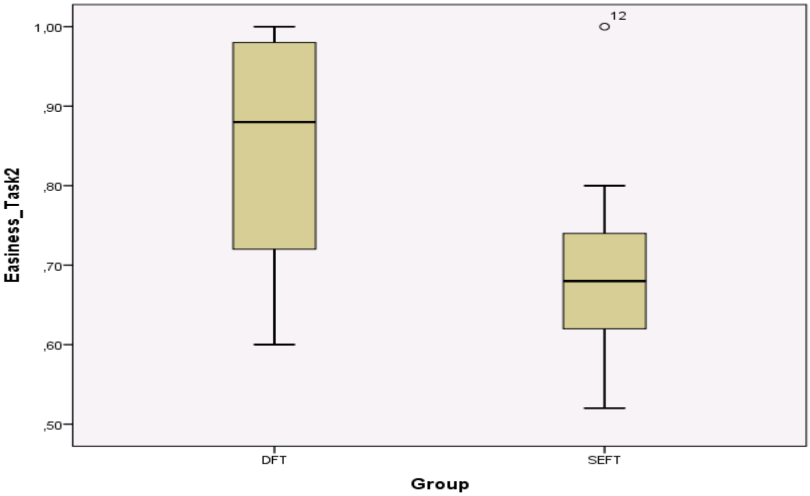

Figure 7 Boxplot: Easiness for task 02

Easiness for Task 03

As shown in Table 11 the easiness score for task 03 has a median of 0,60 in DFT-Group and 0,68 in SEFT group. A detailed view of T03.02, T03.08, T03.09, T03.10 and T03.11 shows us that the difference is on answers for T03.08 where the result is better for the SEFT-Group. A hypothesis test does not confirm the results of median calculation with an acceptable confidence level.

Descriptives				
	Group		Statistic	Std. Error
Easiness_Task3	DFT	Mean	,6743	,06260
		Median	,6000	
		Variance	,027	
		Std. Deviation	,16562	
		Minimum	,48	
		Maximum	,92	
	SEFT	Mean	,6114	,07986

		Median	,6800	
		Variance	,045	
		Std. Deviation	,21130	
		Minimum	,20	
		Maximum	,84	

Table 11 Descriptive statistics: Easiness for Task 03

Easiness scores in task 03 were normaly distributed for both groups as assessed by the Shapiro-Wilk's test (Sig.= 0,404 for DFT group and Sig. = 0,186 for SEFT group). Because of the outliers (Figure 8) we could not compare the mean. Therefore we only qualitatively compare the median obtained previously and see that the median for the SEFT group is slightly better than the one for the DFT group.

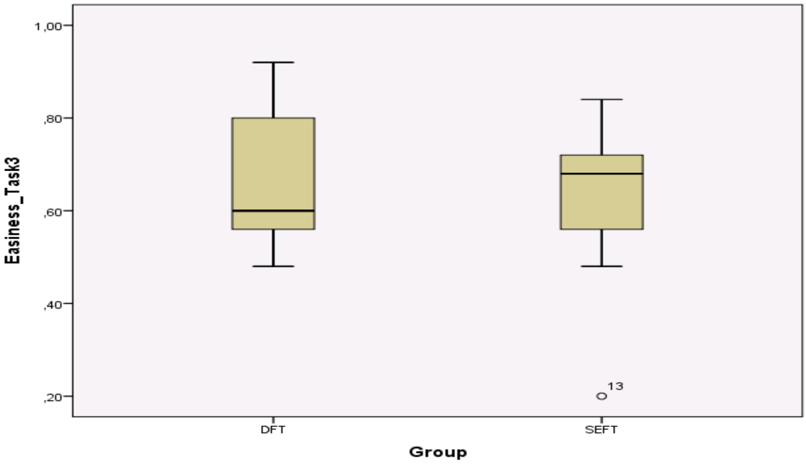

Figure 8 Boxplot: Easiness for task 03

4.3.3 Understandability

Understandability measures the effort needed to understand the models built with the technique in relation to the failure logic. Understandability score for task X is calculated from the answer of TX.05, TX.06 and TX07.

Understandability scores decrease with complexity only in the DFT group. It decreases from the 1st to the 2nd task and increases from the 2nd to the 3rd task. For the 1st task the score is very high for both groups (Median = 0,9 for DFT group and 0,86 for SEFT group). In the 2nd task the score for the DFT group stays at a good level (Median = 0,9) when the score for the SEFT group decreases to Median=0,66. The situation observed from the 1st to the 2nd task for the DFT group is same from the 2nd to the 3rd task in which the score for the DFT group stays at a good level

(Median=0,86). Surprisingly the score for the SEFT group increases from the 2nd to the 3rd task, going from Median=0,66 to Median = 0,8.

For the 1st and 2nd task Understandability scores were better for the DFT group, but for the 3rd task they were similar.

Statistical significance for our comparison could not be achieved for all quantitative computations.

Following we present in detail the analysis of Easiness for each task.

Understandability for Task 01

As shown in Table 12 the understandability score for task 01 has a median of 0,9333 in DFT-Group and 0,8667 in SEFT group. A detailed view of T01.05, T01.06 and T01.07 shows us that the difference is on answers for T01.05 and T01.06 where the results are better for the DFT-Group. A hypothesis test confirms the results of median calculation for both groups with an acceptable confidence level (0,017 for DFT group and 0,027 for SEFT group).

Descriptives				
	Group		Statistic	Std. Error
Understandability_Task1	DFT	Mean	,8952	,04330
		Median	,9333	
		Variance	,013	
		Std. Deviation	,11455	
		Minimum	,73	
		Maximum	1,00	
	SEFT	Mean	,8190	,05767
		Median	,8667	
		Variance	,023	
		Std. Deviation	,15258	
		Minimum	,60	
		Maximum	1,00	

Table 12 Descriptive Statistics: Understandability for Task 01

Understandability scores in task 01 were normaly distributed for both groups as assessed by the Shapiro-Wilk's test (Sig.= 0,073 for DFT group and Sig. = 0,461 for SEFT group). There were no outliers in the data, as assessed by inspection of the boxplot (Figure 9). Therefore we perform an independent-samples t-test to compare the mean obtained for both groups. Understandability score in task 01 for DFT-Group (Mean = 0,8952 ± 0,114) was 0,76 higher than understandability score in task 01 for SEFT-Group (Mean = 0,8190 ± 0,152). The mean difference (0,76) was not satistically significant (p=0,312).

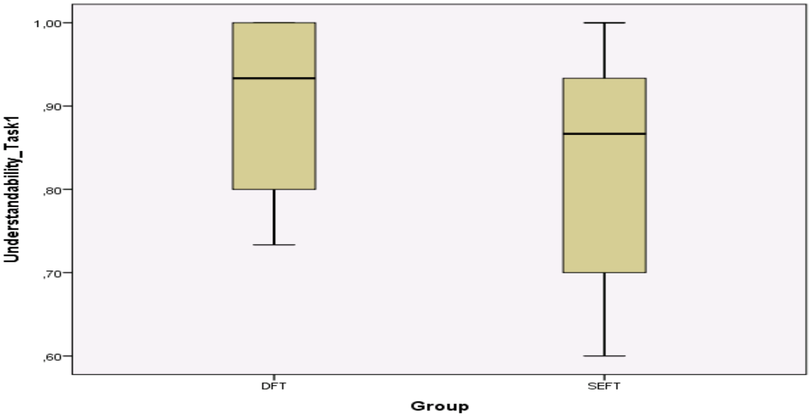

Figure 9 Boxplot: Understandability for Task 01

Understandability for task 02

As shown in Table 13 the understandability score for task 02 has a median of 0,9333 in DFT-Group and 0,6667 in SEFT group. A detailed view of T02.05, T02.06 and T02.07 shows us that the difference is on answers for T01.07 where the results are better for the DFT-Group. A hypothesis test confirms the results of median calculation for both groups with an acceptable confidence level (0,027 for DFT group and 0,016 for SEFT group).

Descriptives				
	Group		Statistic	Std. Error
Understandability_Task2	DFT	Mean	,8571	,05714
		Median	,9333	
		Variance	,023	
		Std. Deviation	,15119	
		Minimum	,60	
		Maximum	1,00	
	SEFT	Mean	,7619	,05009
		Median	,6667	
		Variance	,018	
		Std. Deviation	,13254	
		Minimum	,67	
		Maximum	1,00	

Table 13 Descriptive Statistics: Understandability for Task 02

Understandability scores in task 02 were not normaly distributed for DFT group as assessed by the Shapiro-Wilk's test (Sig.= 0,262). Therefore we perform a median test to compare the median obtained previously. The median test retains the null hypothesis. We couldn't have a statistical significant difference between both median. But a qualitative analysis shows us that the median for the DFT group is better than in the SEFT group (Figure 10).

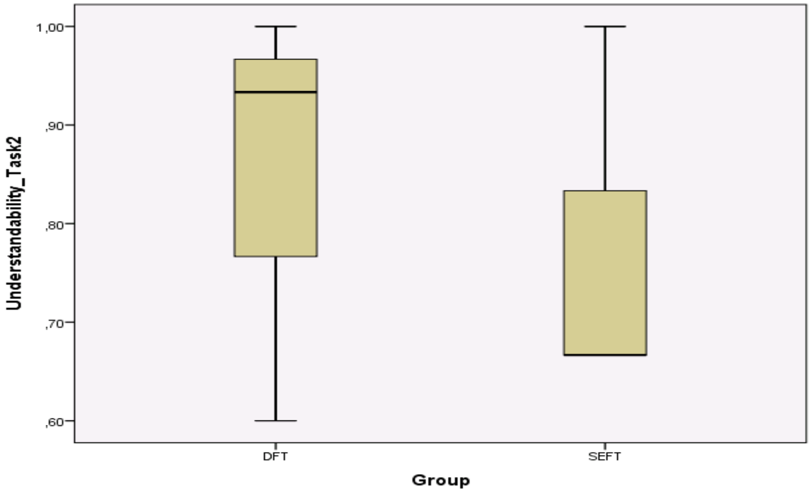

Figure 10 Boxplot: Understandability for Task 02

Understandability for task 03

As shown in Table 14 the understandability score for task 03 has a median of 0,8667 in DFT-Group and 0,8000 in SEFT group. A hypothesis test confirms the results of median calculation only for the DFT group with an acceptable confidence level (0,017).

Descriptives				
	Group		Statistic	Std. Error
Understandability_Task3	DFT	Mean	,8190	,05767
		Median	,8667	
		Variance	,023	
		Std. Deviation	,15258	
		Minimum	,67	
		Maximum	1,00	
	SEFT	Mean	,6667	,08729
		Median	,8000	
		Variance	,053	

		Std. Deviation	,23094	
		Minimum	,20	
		Maximum	,87	

Table 14 Descriptive Statistics: Understandability for Task 03

Understandability scores in task 03 were not normaly distributed for both groups as assessed by the Shapiro-Wilk's test (Sig.= 0,055 for DFT group and Sig. = 0,036 for SEFT group). Therefore we perform a median test to compare the median obtained previously. The median test retain the null hypothesis: the median of understandability score in both groups are the same (Figure 11).

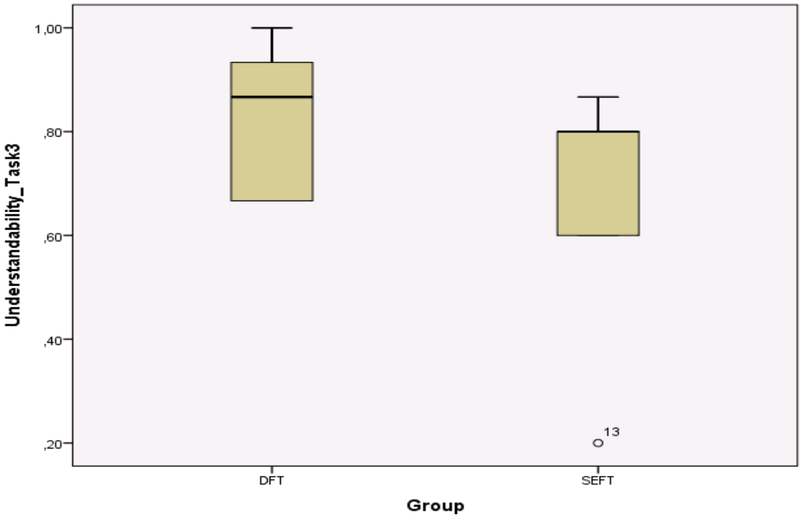

Figure 11 Boxplot: Understandability for Task 03

4.3.4 Time needed

Time needed measures the time needed for building the models.

As shown in Table 15 the time needed has a mean of:

- 7,86 in DFT-Group and 12,86 in SEFT group for Task 01
- 9,14 in DFT-Group and 21,14 in SEFT group for Task 02
- 18 in DFT-Group and 27,43 in SEFT group for Task 03

Descriptives				
	Group		Statistic	Std. Error
Time_Needed_Task1	DFT	Mean	7,86	2,355
		Median	5,00	
		Variance	38,810	
		Std. Deviation	6,230	

		Minimum	2	
		Maximum	20	
	SEFT	Mean	12,86	2,577
		Median	14,00	
		Variance	46,476	
		Std. Deviation	6,817	
		Minimum	3	
		Maximum	23	
Time_Needed_Task2	DFT	Mean	9,14	2,272
		Median	7,00	
		Variance	36,143	
		Std. Deviation	6,012	
		Minimum	4	
		Maximum	20	
	SEFT	Mean	21,14	1,908
		Median	22,00	
		Variance	25,476	
		Std. Deviation	5,047	
		Minimum	12	
		Maximum	26	
Time_Needed_Task3	DFT	Mean	18,00	3,251
		Median	16,00	
		Variance	74,000	
		Std. Deviation	8,602	
		Minimum	9	
		Maximum	30	
	SEFT	Mean	27,43	3,798
		Median	23,00	
		Variance	100,952	
		Std. Deviation	10,048	
		Minimum	15	
		Maximum	40	

Table 15 Descriptive Statistics: Time needed for completing tasks

Time needed in task 01, 02 and 03 were normaly distributed for both groups as assessed by the Shapiro-Wilk's test. We perform an independent-samples t-test to compare the mean of time needed obtained for both groups in each task:

- For task 01: Time needed in the DFT-Group (Mean = 7,86 ± 6,23) was 5,00 lower than time needed in the SEFT-Group (Mean =

12,86 ± 6,81). The mean difference (5,00) was not satistically significant (p=0,178).

- For task 02: Time needed in the DFT-Group (Mean = 9,14 ± 6,01) was 12,00 lower than time needed in the SEFT-Group (Mean = 21,14 ± 5,04). The mean difference (12,00) was satistically significant (p=0,002).
- For task 03: Time needed in the DFT-Group (Mean = 18,00 ± 8,602) was 9,42 lower than time needed in the SEFT-Group (Mean = 27,43 ± 10,04). The mean difference (9,42) was not satistically significant (p=0,084).

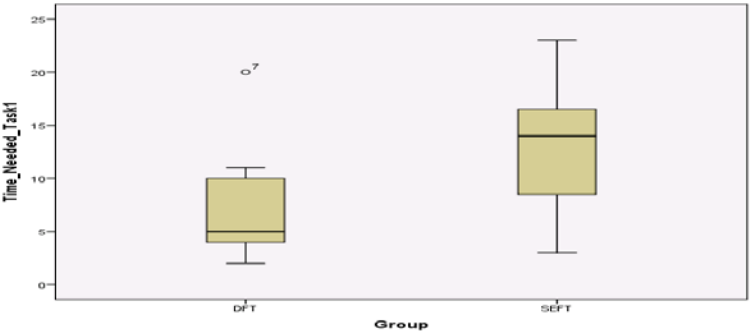

Figure 12 Boxplot: Time needed for task 01

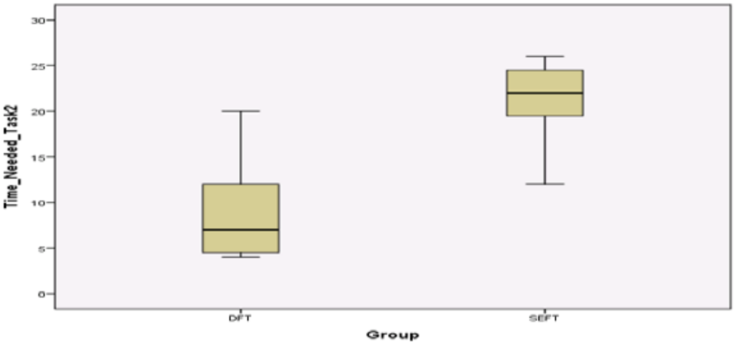

Figure 13 Boxplot: Time needed for task 02

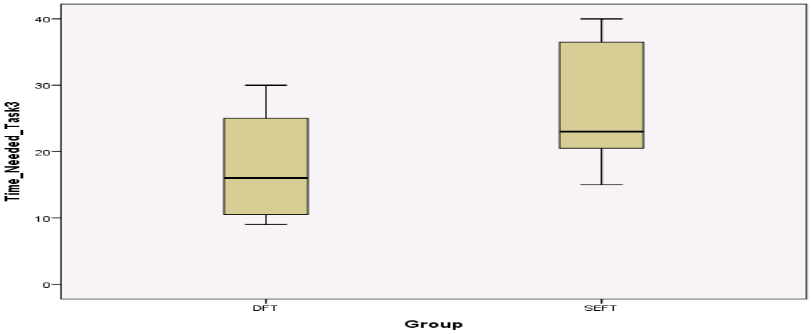

Figure 14 Boxplot: Time needed for task 01

4.3.5 Quality of produced models

Quality of produced models measures how good produced models are in term of the appropriateness of choosen elements for building the tree.

As shown in the time needed has a mean of:

- 6,71 in DFT-Group and 7,42 in SEFT group for Task 01
- 4,85 in DFT-Group and 5,85 in SEFT group for Task 02
- 3,21 in DFT-Group and 3,92 in SEFT group for Task 03

Descriptives				
	Group		Statistic	Std. Error
Model_Quality_Task1	DFT	Mean	6,7143	,77810
		Median	6,0000	
		Variance	4,238	
		Std. Deviation	2,05866	
		Minimum	4,00	
		Maximum	10,00	
	SEFT	Mean	7,4286	1,57143
		Median	10,0000	
		Variance	17,286	
		Std. Deviation	4,15761	
		Minimum	,00	
		Maximum	10,00	
Model_Quality_Task2	DFT	Mean	4,8571	,63353
		Median	6,0000	
		Variance	2,810	
		Std. Deviation	1,67616	
		Minimum	2,00	
		Maximum	6,00	
	SEFT	Mean	5,8571	,73771

		Median	6,0000	
		Variance	3,810	
		Std. Deviation	1,95180	
		Minimum	3,00	
		Maximum	8,00	
Model_Quality_Task3	DFT	Mean	3,2143	,66240
		Median	3,5000	
		Variance	3,071	
		Std. Deviation	1,75255	
		Minimum	,00	
		Maximum	5,00	
	SEFT	Mean	3,9286	,79003
		Median	3,0000	
		Variance	4,369	
		Std. Deviation	2,09023	
		Minimum	1,50	
		Maximum	7,00	

Table 16: Descriptive statistics: Quality of produced models

The quality of produced models in task 01, 02 and 03 were normaly distributed for at least one group as assessed by the Shapiro-Wilk’s test. Therefore we perform an independent-samples t-test to compare the mean obtained for both groups in each task:

- For task 01: The quality for produced model in the DFT-Group (Mean = 6,71 ± 2,05) was 0,71 lower than the quality of produced models in the SEFT-Group (Mean = 7,42 ± 4,15). The mean difference (0,71) was not statistically significant (p=0,691).
- For task 02: The quality for produced model in the DFT-Group (Mean = 4,85 ± 1,67) was 1,00 lower than the quality of produced models in the SEFT-Group (Mean = 5,85 ± 1,95). The mean difference (1,00) was not statistically significant (p=0,324).
- For task 03: The quality for produced model in the DFT-Group (Mean = 3,21 ± 1,75) was 0,71 lower than the quality of produced models in the SEFT-Group (Mean = 3,92 ± 2,09). The mean difference (0,71) was not statistically significant (p=0,502).
-

4.3.6 Effort expectancy

As shown in Table 17Table 7 the effort expectancy score has a median of 0,8 in DFT-Group and 0,8 in SEFT group. A hypothesis test confirms the results of median calculation only for the DFT-Group with an acceptable confidence level (0,018)

Descriptives

	Group		Statistic	Std. Error
Effort_Expectancy	DFT	Mean	,8357	,05084
		Median	,8000	
		Variance	,018	
		Std. Deviation	,13452	
		Minimum	,65	
		Maximum	1,00	
	SEFT	Mean	,7500	,06986
		Median	,8000	
		Variance	,034	
		Std. Deviation	,18484	
		Minimum	,45	
		Maximum	,95	

Table 17 Descriptive Analysis: Effort Expectancy

Effort expectancy scores were normaly distributed for both groups as assessed by the Shapiro-Wilk's test (Sig.= 0,424 for DFT group and Sig. = 0,390 for SEFT group). There were no outliers in the data, as assessed by inspection of the boxplot (Figure 15). Therefore we perform an independent-samples t-test to compare the mean obtained for both groups. Effort expectancy scores for DFT-Group (Mean = 0,8357 ± 0,134) was 0,085 higher than Effort expectancy scores for SEFT-Group (Mean = 0,75 ± 0,184). The mean difference (0,085) was not statistically significant (p=0,341).

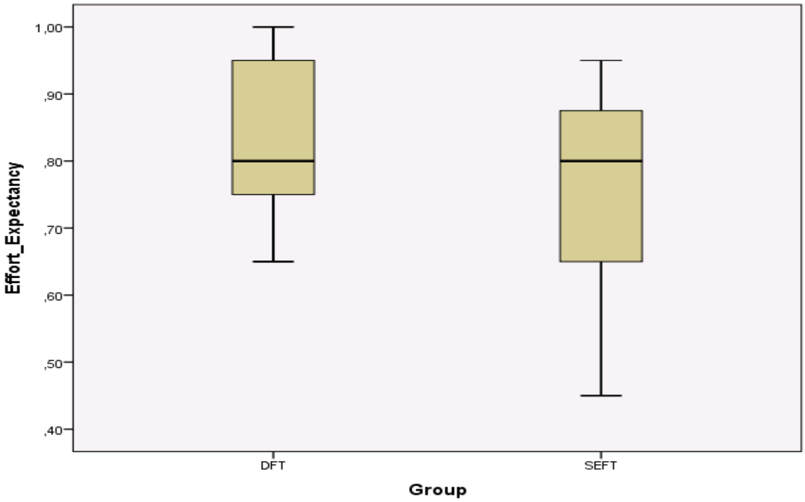

Figure 15 Boxplot: Effort expectancy

4.3.7 Attitude toward using the methodology

Attitude toward using the methodology measures the overall affective reaction to using the technique.

As shown in Table 18Table 7 the attitude with technology score has a median of 0,75 in DFT-Group and 0,75 in SEFT group. A detailed view of D.05, D.06, D.07 and D.08 shows us that the difference is on answers for D.06 where the results a better for the DFT-Group. A hypothesis test confirms the results of median calculation only for the DFT-Group with an acceptable confidence level (0,018)

Descriptives				
	Group		Statistic	Std. Error
Attitude_with_Technology	DFT	Mean	,7714	,04345
		Median	,7500	
		Variance	,013	
		Std. Deviation	,11495	
		Minimum	,65	
		Maximum	1,00	
	SEFT	Mean	,7143	,04592
		Median	,7500	
		Variance	,015	
		Std. Deviation	,12150	
		Minimum	,55	
		Maximum	,90	

Table 18 Descriptive Statistics: Atiitude with the technology

Attitude with technology scores were normaly distributed for both groups as assessed by the Shapiro-Wilk's test (Sig.= 0,182 for DFT group and Sig. = 0,883 for SEFT group). Therefore we perform an independent-samples t-test to compare the mean obtained for both groups. Attitude with technology scores for DFT-Group (Mean = 0,77 ± 0,114) was 0,057 higher than Effort expectancy scores for SEFT-Group (Mean = 0,71 ± 0,121). The mean difference (0,057) was not statistically significant (p=0,384).

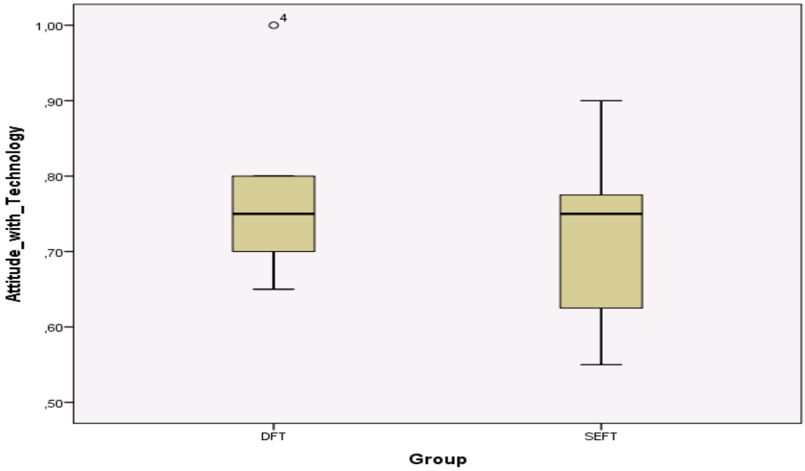

Figure 16 Boxplot: Attitude with technology

4.3.8 Self efficacy

Self efficacy measures the degree to which the subject believes that they will better perform if they have some help.

As shown in Table 19 the self efficacy score has a median of 0,75 in DFT-Group and 0,70 in SEFT group. A detailed view of D.09, D.10, D.11 and D.12 shows us that the difference is on answers for D.09 where the results a better for the DFT-Group. A hypothesis test confirms the results of median calculation only for the DFT-Group with an acceptable confidence level (0,027)

Descriptives				
	Group		Statistic	Std. Error
Self_Efficacy	DFT	Mean	,7417	,07350
		Median	,7500	
		Variance	,032	
		Std. Deviation	,18005	
		Minimum	,50	
		Maximum	1,00	
	SEFT	Mean	,7071	,02974
		Median	,7000	
		Variance	,006	
		Std. Deviation	,07868	
		Minimum	,60	
		Maximum	,80	

Table 19: Descriptive Statistics: Self efficacy

Self efficacy scores were normaly distributed for both groups as assessed by the Shapiro-Wilk's test (Sig.= 0,988 for DFT group and Sig. = 0,420 for SEFT group). There were no outliers in the data, as assessed by inspection of the boxplot (Figure 17). Therefore we perform an independent-samples t-test to compare the mean obtained for both groups. Self efficacy scores for DFT-Group (Mean = 0,74 ± 0,18) was 0,034 higher than Effort expectancy scores for SEFT-Group (Mean = 0,70 ± 0,078). The mean difference (0,034) was not statistically significant (p=0,654).

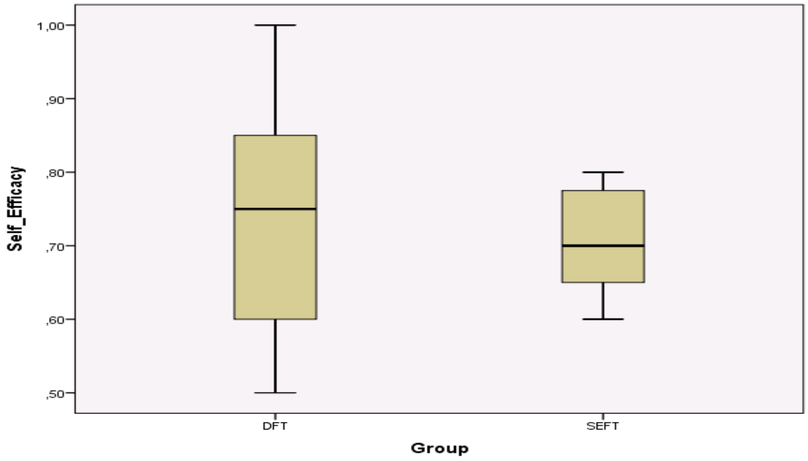

Figure 17 Boxplot: Self efficacy

4.3.9 Performance expectancy

Performance expectancy measures the degree to which the subject believes that using the technique will help him to attain gains in job performance. This metric was only valid for people who are professionnaly active (researchers or assistant researchers).

As shown in Table 20 the performance expectancy score has a median of 0,725 in DFT-Group and 0,60 in SEFT group. A detailed view of D.13, D.14, D.15 and D.16 shows us that the difference is on answers for D.13, D.14 and D15 where the results are better for the DFT-Group. A hypothesis test confirms the results of median calculation only for the DFT-Group with an acceptable confidence level (0,043)

Descriptives[a]

	Group		Statistic	Std. Error
Performance_Expectancy	DFT	Mean	,7583	,05974
		Median	,7250	

		Variance	,021	
		Std. Deviation	,14634	
		Minimum	,60	
		Maximum	1,00	
	SEFT	Mean	,6333	,06009
		Median	,6000	
		Variance	,011	
		Std. Deviation	,10408	
		Minimum	,55	
		Maximum	,75	

Table 20 Descriptive Statistic: Performance expectancy

Performance expectancy scores were normaly distributed for both groups as assessed by the Shapiro-Wilk's test (Sig.= 0,682 for DFT group and Sig. = 0,463 for SEFT group). There were no outliers in the data, as assessed by inspection of the boxplot (Figure 18). Therefore we perform an independent-samples t-test to compare the mean obtained for both groups. Performance expectancy scores for DFT-Group (Mean = 0,75 ± 0,14) was 0,125 higher than Performance expectancy scores for SEFT-Group (Mean = 0,63 ± 0,1). The mean difference (0,125) was not statistically significant (p=0,234).

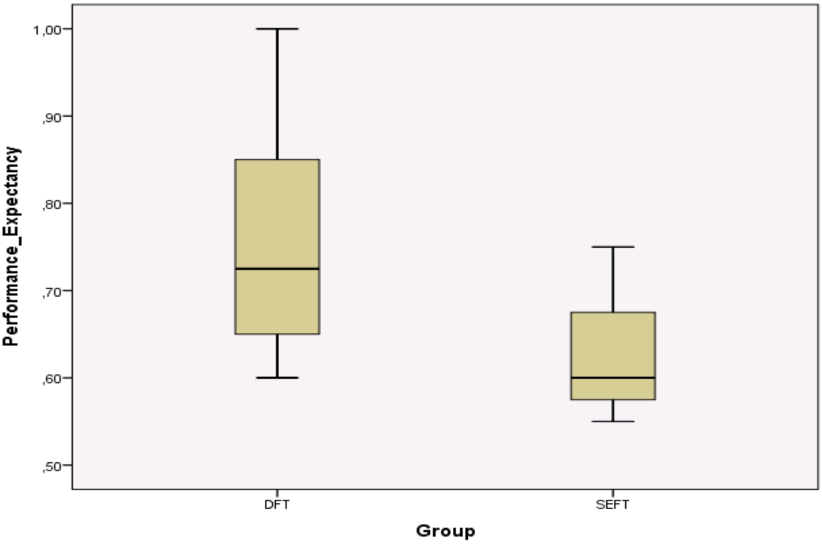

Figure 18 Boxplot: Performance expectancy

4.3.10 Comments from subjects

A coding analysis [coding analysis] was performed to analyze comments made by subjects and results are shown in ref table

DFT	SEFT
Which advantages do you see using DFTs/SEFTs?	
Completeness: more gates (priority, order of events, dependence between events) than FT, no possibility to represent sequence of events **Model accuracy:** accurate models	**Completeness:** better expressiveness **Understandability:** easy to use and understand, graphical representation
Which disadvantages do you see in using DFTs/SEFTs?	
Time modeling: not possible **Understandability:** not easy to understand, not easy to understand notation, big models difficult to comprehend **Completeness:** no possibility to represent sequence of events	**Understandability:** mixing gates and events makes it difficult to understand, redundant information **Expressiveness:** It is hard to represent looping conditions **Usability:** difficult to use as petrinets

Table 21 Comments from subjects

4.4 Analysis Summary

Table 22 shows a summary of our analysis. In the column hypothesis check, a + (resp. -) shows the satisfaction (resp. non satisfaction) of the main hypothesis (SEFT obtains better results than DFT).

Metric		Mean / Median		Hypothesis Check
		SEFT	DFT	
Completeness	Task 01	,9333	,9333	-
	Task 02	,6667	,8667	-
	Task 03	,6667	,7333	-
Easiness	Task 01	,9200	,9600	-
	Task 02	,6800	,8800	-
	Task 03	,6800	,6000	+
Understandability	Task 01	,8667	,9333	-
	Task 02	,6667	,9333	-
	Task 03	,8000	,8667	-
Time needed (min)	Task 01	12,86	7,86	-
	Task 02	21,14	9,14	-
	Task 03	27,43	18,00	-
Quality of produced models	Task 01	7,4286	6,7143	+
	Task 02	5,8571	4,8571	+
	Task 03	3,9286	3,2143	+
Effort Expectancy		,8000	,8000	-
Attitude toward using the method		,7500	,7500	-
Self- Efficacy		,7000	,7500	-

Performance Expectancy	,6000	,7250	-

Table 22 Analysis summary

5 Discussion

SEFT and DFT techniques were analyzed for obtain measures for metrics in order to answer two questions:

- How appropriate were the two models?
- How easy was the technique to use?

To answer the first question the metrics used were completeness, understandability, attitude and quality of produced models. To obtain measures for these metrics, participants of a controlled experiment were asked questions with respect to some tasks such as changing the SEFTs or DFTs when a new component or behavior was introduced into the system. The participants then had to judge how well they were able to accomplish these tasks with the methodology (SEFT or DFT) at hand. Based the answers given by the participants, we obtained the following results for the two techniques:

- SEFTs and DFTs help in achieving completeness of models by being able to transfer system description, identifying appropriate gates and co-relate it to the failure model(SEFT/DFT) for tasks that involved adding new failure modes or changing the failure behavior. But for tasks that involved modeling using more advanced gates (those other than the simple and/or gates) the measure for completeness of both techniques decreases.
- We measured understandability by measuring how easy it was to keep an overview of the entire failure logic and co-related the system model to the failure model. We found that DFTs were more understandable than SEFTs irrespective of the complexity/nature of the tasks, but SEFTs proved to be difficult to understand while changing the failure behavior by adding new failure modes and logic to components.
- Both DFT and SEFTs enjoy a good attitude towards their usage where the participants think that using them make their work more fun and interesting. For both groups the quality of produced models decreases with complexity, but the SEFT group always produces better safety models than the DFT group.
- We noticed that although subjects of the DFT group believe that they were more able (completeness and understandability) to perform their respective tasks than subjects of the SEFT group, the feeling about the technology was the same (attitude with technology) and subjects of the SEFT group even produce better models (quality of produced models) than subject of the DFT group.

Hence we can say that for both techniques, the appropriateness depends on the nature and complexity of the task that had to be accomplished. Each technique has its own inherent advantage and disadvantage. For example, while it is very easy to make local changes in an SEFT, it may be challenging to keep in mind the overall failure scenario. DFTs on the other hand give a quick overview of the failure scenario, but it may be a challenge to locate the point of change in the DFT.

Along with the above metrics, we also measured the easiness, time needed, self-efficacy, effort and performance expectancy in order to answer the second question.

As expected, the easiness decreased, the time needed increased as the complexity of tasks increased for both techniques. Although both techniques score high on self-efficacy, it was found that most users felt they could do better if they had some assistance and guidance. Effort expectancy scores shows us that the ease of use associated with both techniques, although very good, are similar. Researchers among subjects of the DFT group think that their chance of improving their job performance by using DFT is quite good. Researchers among subjects of the SEFT group are more pessimists with a score close to our threshold value.

The measures obtained for the metrics easiness, understandability and time needed show that DFTs are easier to use as compared to SEFTs. Both techniques had similar measures for other metrics used to answer the second question.

6 Threats to validity

In this section we discussed how validity threats [15] were avoided.

6.1.1 Conclusion validity

Due to the low number of subjects (2 groups of 7 subjects each) we couldn’t avoid the low power of performed statistical tests. Nevertheless we could confirm some conclusion of statistical tests based on comments and feedbacks provided by subjects. Subjects of both groups have similar background and knowledge about safety analysis (see Table 1 and Table 2).

Before performing test preconditions (normality, independence of variables …) were checked to make sure that they are satisfied.

To get reliable measures questionnaires were checked by an expert on empirical studies.

6.1.2 Internal validity

Subjects were trained with techniques before experimentation and based on feedbacks we concluded that the group performing DFT has a better understanding of the technique than the group performing SEFT. The main reason for this was that DFT uses notation closed to FTA which is taught in their safety analysis lecture.

To avoid any learning groups were trained and applied their respective techniques in different rooms.

6.1.3 Construct validity

The experiment’s goal was refined into clearly defined metrics and measures to avoid misunderstandings.
Different tasks were proposed to subjects of both groups for avoiding mono-operation bias.
Subjects only apply one technique depending on the group they belong to.
Subjects were not aware of the hypothesis to be tested or measures to be taken.

6.1.4 External validity

We tried to have subjects closed enough to industrial setting by selecting graduate students who have some knowledge of safety analysis. Both groups also contain researchers with some years experience on safety analysis.

The proposed system was derived from a system used in a living lab (close enough to real setting).

7 Conclusion

In order to analyze the applicability and efficiency of SEFTs and DFTs on modeling safety aspects of ambient assisted living systems, we performed a controlled experiment where subjects have to apply these techniques on an AAL system and provide their feedback on using these techniques.

The experiment was conducted with students and researchers of the TU Kaiserslautern and consisted of two parts: a training session where subjects were trained on using the techniques and the main experimental session where they used the techniques for modeling different safety aspects of an ambient assisted living system.

Results of the experiment show that students found DFTs more easy and effective to use, althought they produce better results (models) with SEFTs.

We could not obtain enough data to statistically support our results and therefore we are planning to replicate the experiment with more subjects.

8 References

[1] T. Kleinberger, M. Becker, E. Ras, A. Holzinger, and P. Müller, "Ambient intelligence in assisted living: enable elderly people to handle future interfaces," in *Proceedings of the 4th international conference on Universal access in human-computer interaction: ambient interaction*, ser. UAHCI'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 103–112.

[2] R. Oppermann and R. Rasher, "Adaptability and adaptivity in learning systems," *Knowledge transfer*, vol. 2, pp. 173–179, 1997.

[3] J. Nehmer, M. Becker, A. Karshmer, and R. Lamm, "Living assistance systems: an ambient intelligence approach," in *Proceedings of the 28th international conference on Software engineering*, ser. ICSE '06. New York, NY, USA: ACM, 2006, pp. 43–50.

[4] S. Spang, "Scrutinizing the impact of security on safety on an communicating vehicle pla-toon." VIERForES Meilensteinbericht zum Arbeitspaket 5.5.2.

[5] R. Schwarz, "Modulare security-modelle zur komponentenbasierten modellierung komplexer eingebetteter systeme," Fraunhofer IESE, Milestone report to Workpackage 6.1.3, 2009.

[6] B. Kaiser, P. Liggesmeyer, and O. Mïckel, "A new component concept for fault trees," in *Proceedings of the 8th Australian Workshop on Safety Critical Systems and Software (SCSS03)*, 2003, pp. 37–46.

[7] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr, "Basic concepts and taxonomy of dependable and secure computing," *IEEE Transactions on Dependable and Secure Computing*, vol. 1, no. 1, pp. 11 – 33, jan.-march 2004.

[8] W. E. Vesely, F. F. Goldberg, N. H. Roberts, and D. F. Haasl, *Fault Tree Handbook*. U.S. Nuclear Regulatory Commission, 1981.

[9] P. Liggesmeyer, *Qualitätssicherung softwareintensiver technischer Systeme*, ser. Forschung in der Softwaretechnik. Spektrum Akademischer Verlag, 2000.

[10] N. G. Leveson, *Safeware - system safety and computers: a guide to preventing accidents and losses caused by technology*. Addison-Wesley, 1995.

[11] W. Vesely, J. Dugan, J. Fragola, Minarick, and J. Railsback, “Fault tree handbook with aerospace applications,” National Aeronautics and Space Administration, Washington, DC, Handbook, 2002.

[12] B. Kaiser, “State event fault trees: a safety and reliability analysis technique for software controlled systems,” Ph.D. dissertation, Technical University of Kaiserslautern, 2006.

[13] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, “User acceptance of information technology: Toward a unified view,” *MIS quarterly*, pp. 425–478, 2003.

[14] V. Venkatesh and H. Bala, “Technology acceptance model 3 and a research agenda on interventions,” *Decision sciences*, vol. 39, no. 2, pp. 273–315, 2008.

[15] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering: an introduction*. Norwell, MA, USA: Kluwer Academic Publishers, 2000.