Supporting Expert Assessment of Argument Structures in Trust Cases

Lukasz Cyra[a*], Janusz Gorski[a†]
[a]Gdansk University of Technology, Gdansk, Poland

Abstract: Trustworthiness of a selected system or service can be justified using the concept of a trust case. Trust case denotes a complete and explicit argument and supporting evidence, which aim at justifying the trustworthiness of a chosen object in its target environment. Trust cases can grow to complex structures which include arguments of various types and evidence of different quality. Two objectives are of particular interest: (1) assessment of the compelling power of the trust case, and (2) communication of the result of such an analysis. To address the above objectives, a mechanism of gathering expert opinions about the value of evidence and about the validity of arguments has been introduced. The mechanism is based on the Dempster-Shaffer model, and is well suited to represent uncertainty resulting from the lack of knowledge of experts and to combine evidence of varying levels of consistency. The paper introduces the Trust-IT framework, which supports development and application of trust cases. Then different types of argument strategies used in trust cases are identified, and adequate rules of aggregation are proposed for each of them. Finally, the results of a set of experiments aiming at assessment of the validity of the rules are presented.

Keywords: Trust Case, Assurance Case, Dempster-Shafer Model, Argument Assessment

1. INTRODUCTION

Arguments are used to justify in an explicit way various qualities of different objects. Application of arguments to justify safety in so called *safety cases* is well established [1]. Recently there is a growing interest in (often less rigorous) arguments which could be used, for instance, to demonstrate security [2], trustworthiness [3, 4], conformity with standards [5, 6], or the adequacy of metrics selected for the chosen measurement objectives [7]. The 'hosting' data structures of such arguments are known as assurance cases, trust cases, conformity cases and so on.

The idea which underpins the development of argument structures is to make expert judgment explicit in order to redirect dependence on judgment to issues on which we can trust this judgment [8]. However, such an approach is not sufficient as argument structures tend to grow excessively. This leads to the necessity of expert assessment of the compelling power of the structures themselves. Although this is a commonly accepted practice, this approach has a significant inherent drawback. The

[*] lukasz.cyra@eti.pg.gda.pl
[†] jango@pg.gda.pl

research in experimental psychology shows that human minds do not deal properly with complex inference based on uncertain sources of knowledge [8], which is common in trust cases.

Therefore, an appropriate mechanism of gathering expert opinions about the value of evidence and about validity of arguments is required. Experts should only be required to express their opinions about the basic elements in argument structures, in which they are sufficiently reliable. Then the mechanism should provide for aggregation of the opinions into an easy to communicate message about the quality of the overall argument.

The objective of this article is to present such an aggregation mechanism. To this end, the paper presents Trust-IT, which is a framework for argument structures development [9]. Then relevant qualities of trust cases, concerning automatic expert opinions aggregation, are discussed and different argumentation strategies are identified. Finally, the aggregation mechanism is presented together with the results of its experimental validation.

The presented method extends and modifies the approach to trust case appraisal proposed [10], which had problems with adapting itself to the types of arguments occurring in trust cases. Consequently, some basic assumptions concerning the aggregation mechanism had to be changed, which resulted in a complete rework of the mechanism of issuing and aggregating expert opinions.

2. TRUST-IT

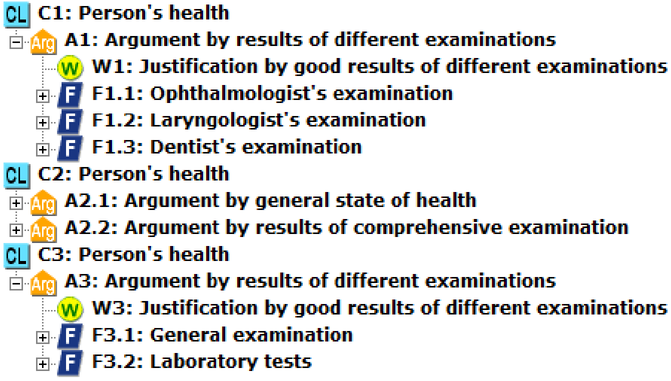
Trust case denotes a complete argument and supporting evidence, which aim at justifying certain qualities of a chosen object. It is developed by making an explicit set of claims about the object and showing how the claims are interrelated and supported by the evidence.

Trust-IT is a framework for the development and application of trust cases [9]. The framework consists of three components: an application, methodological and tool component. The application component explains possible usage scenarios for trust cases. The methodological component explains how to develop and maintain trust cases. The tool component provides for full scale exploitation of the two other components.

The framework defines a language to express trust cases. The language distinguishes different components which represent different elements of an argument structure. The whole structure has a tree like form.

The basic building block of a trust case consists of a conclusion to be justified i.e. a *claim* (denoted CL), the *premises* which are referred to in order to justify the conclusion and the justification which explains why the conclusion results from the premises. Such a block is called an *argument*. Each of the claims in a trust case is supported by an *argument strategy* (denoted Arg) which contains the basic idea how to derive a conclusion from premises. In case of *counter-arguments* it includes the idea of rebuttal of the claim. The justification of the inference from the premises to the conclusion is represented as a node of type *warrant* (denoted W). A premise can be of three different types: it can represent an *assumption* (denoted As), in which case the premise is accepted without further justification; it can be a more specific claim which is justified further; or it can represent a *fact* (denoted F), which is obviously true or otherwise is supported by some evidence. The evidence is provided in additional documents which are pointed to by nodes of type *reference* (denoted Ref).

Figure 1: Trust Case Examples



CL C1: Person's health
- Arg A1: Argument by results of different examinations
  - W W1: Justification by good results of different examinations
  - F F1.1: Ophthalmologist's examination
  - F F1.2: Laryngologist's examination
  - F F1.3: Dentist's examination

CL C2: Person's health
- Arg A2.1: Argument by general state of health
- Arg A2.2: Argument by results of comprehensive examination

CL C3: Person's health
- Arg A3: Argument by results of different examinations
  - W W3: Justification by good results of different examinations
  - F F3.1: General examination
  - F F3.2: Laboratory tests

In figure 1 three example claims are presented. Their justifications are very simple, nevertheless show the idea of building arguments to demonstrate *C1*, *C2* and *C3* which state that a person is healthy.

Claim *C1* is argued on the basis of good results of different medical examinations. The argument strategy is stated in *A1* and explained in detail in warrant *W1*. The argument is based on three premises: facts *F1.1*, *F1.2* and *F1.3*, each of which refers to the results of examinations performed by a different specialist (this reference is not shown in figure 1).

Claim *C2* is argued in two alternative ways, which is represented by two argument strategies: *A2.1* and *A2.2*. The former is based on evidence related to general state of health and the latter refers to results of comprehensive examination.

Claim *C3* (similarly to *C1*) demonstrates health by referring to results of different examinations (argument strategy *A3* and warrant *W3*). However, in this case it refers to different facts: the results of general examination (fact *F3.1*) and laboratory tests (fact *F3.2*).

The question related to the arguments presented in figure 1 is: 'How well do they support the related claims?'. The appraisal mechanism proposed in this paper helps to answer this question by calculating the support given to a claim based on the expert assessments of the basic elements of the argument (i.e. assumptions, facts and warrants).

3. ARGUMENT TYPES

Before introducing the appraisal mechanism we first introduce some assumptions which have influence on the selection of aggregation rules:

- Conclusions and premises of arguments are sentences and we want to assess their acceptability and the confidence in this judgment despite the role a given sentence plays in the trust case (is it a claim, fact or assumption). There are no reasons that the appraisal mechanism should treat claims, facts and assumptions differently.
- Trust cases are represented as trees and the appraisal mechanism can start from the assessment of the most detailed premises (facts and assumptions being the leaves of the tree) and then recursively traverse the tree inferring the assessments of the resulting conclusions.
- A claim in a trust case can be supported by different argument strategies and/or counter-arguments. Therefore, the appraisal mechanism should provide means of aggregating the assessments resulting from different argument strategies and dealing with possible contradictions.
- The support given to a conclusion by a related warrant and premises differs depending on the inference rule used in the warrant. Therefore, the appraisal mechanism should aggregate assessments in a warrant dependent way.

The last point requires more detailed explanation. We distinguish two main types of inference rules occurring in trust cases:

Type 1: rules for which the falsification of a single premise leads to the rebuttal of the conclusion or to the rejection of the whole inference (because nothing can be inferred about the conclusion).

Type 2: rules for which the falsification of one of the premises decreases, but not nullifies, the support for the conclusion. If the remaining premises are accepted, the conclusion can still be attained.

Most of arguments which do not comply with either Type 1 or Type 2 can be represented as a combination of Type 1 and Type 2 arguments.

Arguments of Type 1 can be further divided into 2 sub-categories:

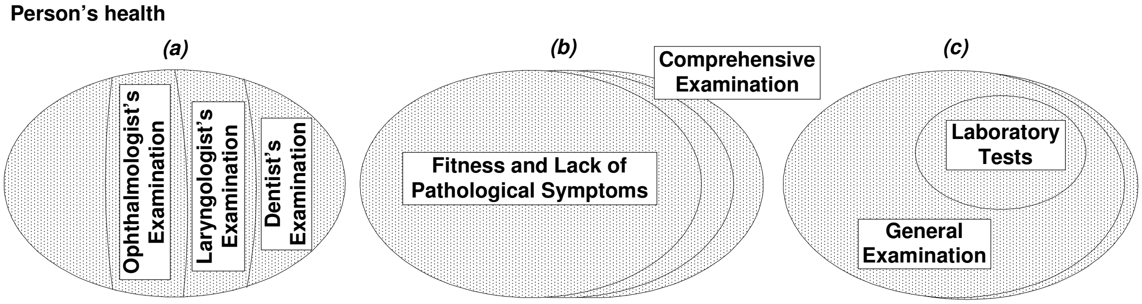
Type 1.1: Acceptance of the premises leads to the acceptance of the conclusion. Falsification of a single premise leads to the rebuttal of the conclusion. This type is called *NSC-argument* (*Necessary and Sufficient Condition list argument).*

Type 1.2: Acceptance of the premises leads to the acceptance of the conclusion. Falsification of a single premise leads to the rejection of the inference. This type of an argument is called *SC-argument* (*Sufficient Condition list argument).*

We can observe that often an argument of Type 1 which does not comply with either NSC- or SC-type can be represented as a combination of NSC- and SC-arguments.

Arguments of Type 2 can be further divided into 3 sub-categories:

Figure 2: Argument Types


Person's health
(a)
Ophthalmologist's Examination
Laryngologist's Examination
Dentist's Examination
(b)
Comprehensive Examination
Fitness and Lack of Pathological Symptoms
(c)
Laboratory Tests
General Examination

Type 2.1: Each of the premises 'covers' part of the conclusion which is represented in figure 2(a), which refers to the argument *A1* in figure 1. The named areas represent parts of the conclusion supported by different facts the argument refers to. The unnamed area represents other aspects which were not covered by the premises. This type of an argument is called *C-argument* (*Complementary argument).*[‡]

Type 2.2: Each of the premises is used in an independent argument and supports the whole conclusion, like in case of arguments *A2.1* and *A2.2* in figure 1. This situation is illustrated in figure 2(b). This type of an argument is called A-argument (*Alternative argument).*

Type 2.3: Each of the premises supports part of the conclusion (not necessarily disjoint) like in the argument *A3* in figure 1. Laboratory tests aim at assessment of some aspects of patient's health and overlap with the results obtained from the general examination (see figure 2(c)). In theory, this case could be treated as a combination of C- and A-arguments. In practice, however, it is not always obvious how to distinguish the overlapping part. We have adopted a pragmatic solution that if an expert assesses the overlapping as insignificant, arguments of Type 2.3 are treated as C-arguments and, if the overlapping is considered significant, the argument is treated as A-arguments.

To summarise, we have identified the need for the following four rules of aggregation:

- to aggregate assessments from multiple argument strategies supporting the same claim (A-arguments),
- to aggregate assessments from premises to a conclusion (NCS-, SC- and C-arguments).

4. THE APPRAISAL MECHANISM

The research presented in this paper has been motivated by the following two objectives related to trust cases:

- Assessment of the compelling power of the trust case - this requires a thorough analysis of trust case contents and can require a considerable expertise and effort.
- Communication of the result of such an assessment to the relevant recipients - this requires that the result of the analysis is passed as a concise and understandable message.

We are addressing the above objectives in the following way:

- Trust case contents are made accessible to a broad range of recipients, called *viewers.*
- Some viewers, called *assessors*, have the right to express their opinion about the trust case contents. It is assumed that an assessor is an expert with sufficient competence to assess the quality of the argument contained in the trust case.

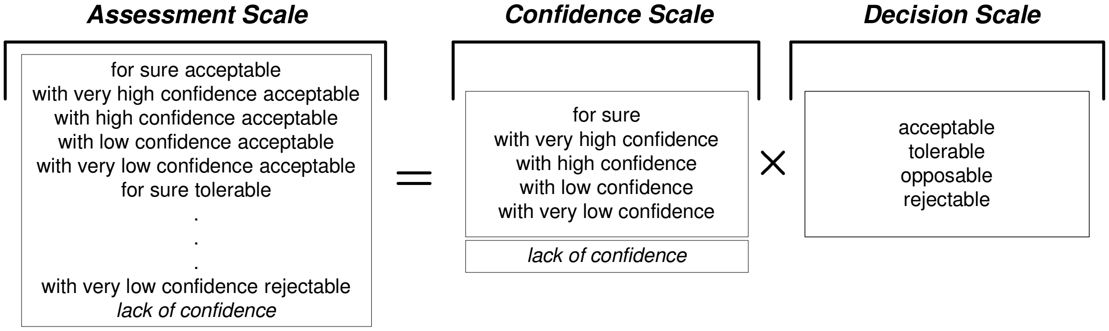
[‡] Another example could be arguing about stability of a table by referring to the facts that it is supported by its legs. Each such fact contributes to the acceptance of the conclusion however, only by referring to at least three such facts we have good support for the conclusion (and the table).

- Assessors can express their opinions about the ‘quality’ of the evidence and about the validity of the arguments included in the trust case.
- The opinions are then aggregated, which results in the opinion about the whole trust case.
- Opinions of different experts are then combined to obtain the sort of ‘trustworthiness index’ of the trust case, which can be communicated to the community of viewers.

The expert assessments and the results of their aggregation are represented in the Dempster-Shafer model [11, 12] by application of a linguistic *Assessment scale* defined in figure 3.

The *Assessment scale* is a product of two more specific scales: *Confidence scale* and *Decision scale*. *Confidence scale* distinguishes six levels of confidence in the subject of assessment. For instance, if the subject of assessment is a statement, a value chosen from the *Confidence scale* expresses the cumulative confidence in acceptance or rejection of this statement (without choosing any of these options). Then, the decision about acceptance (rejection) is expressed referring to the *Decision Scale*.

Figure 3: Trust Assessment Scale



Referring to the Dempster-Shafer’s functions [11, 12]:

$$< Bel(i), Pl(i) > \in [0,1] \times [0,1] \qquad (1)$$

where:

- *i* is a considered statement,
- *Bel(i)* is the *belief* function representing the amount of belief that directly supports *i*,
- *Pl(i)* is a *plausibility* function representing the upper bound on the belief that can be gained by adding new evidence (because there is so much evidence that contradicts it),

we can formally represent *confidence* as:

$$Conf(i) = Bel(i) + 1 - Pl(i) \qquad (2)$$

where *Conf(i)* is a numerical representation of confidence estimation. So:

$$Conf(i) \in [0,1] \qquad (3)$$

and it is taken that *Conf(i) = 0* for *‘lack of confidence’*, and *Conf(i) = 1* for *‘for sure’*.

Decision scale distinguishes four levels to express the ratio between belief (acceptance of a statement) and the overall confidence in the statement (without distinguishing if we want it to be accepted or rejected).

Using Dempster-Shafer’s functions we can formally represent a decision as:

$$Dec(i) = \begin{cases} Bel(i)/(Bel(i) + 1 - Pl(i)) & Bel(i) + 1 - Pl(i) \neq 0 \\ 1 & Bel(i) + 1 - Pl(i) = 0 \end{cases} \qquad (4)$$

where *Dec(i)* is a numerical representation of acceptability estimation. So:

$$Dec(i) \in [0,1] \qquad (5)$$

and *Dec(i) = 0* for *‘rejectable’*, and *Dec(i) = 1* for *‘acceptable’*.

Two scales together provide for expressing both, the attitude towards acceptance or rejection of a statement and the confidence in this decision.

If a considered statement represents a fact of a trust case which is supported by some evidence, then by choosing from *Confidence scale* we express how certain we are (based on the evidence) that the fact should be accepted or rejected. Moving down in the scale we admit more uncertainty about this judgment, whereas moving up we reduce the uncertainty about it. By choosing from *Decision scale* we express our attitude towards acceptance (rejection) of the fact. Higher in the scale we are more towards the acceptance while moving down we are more towards rejection. Thus, by choosing *'for sure'* from *Confidence scale* and *'tolerable'* from *Decision scale* we express that there is no uncertainty that there are reasons to accept and simultaneously there are reasons to reject the fact but in our assessment the reasons towards acceptance are stronger and therefore we can tolerate this fact. If *'with low confidence'* and *'tolerable'* were selected this would mean that the fact is tolerable (there are more reasons to accept than to reject) but there is also much uncertainty where we cannot assess either for or against the fact.

At the bottom of *Confidence scale* there is *'lack of confidence'* value which represents complete indifference with respect to confidence. Therefore, the acceptability assessment does not matter in this case.

The assessment of a statement proceeds as follows. If the assessor can not find any reasons to accept or reject the statement, she/he chooses *'lack of confidence'*. Otherwise, the available evidence (or common knowledge) that supports the judgment (both positive and negative) is assessed using the *Confidence scale* and the (relative) support for acceptability is assessed using the *Decision scale*.

For instance:

- If the available evidence leaves little room for uncertainty (*'with high confidence'*) and all evidence supports acceptability of a premise (*'acceptable'*), then the result is *'with high confidence acceptable'*.
- If premises are irrelevant to a conclusion, then obviously the assessment of a warrant is *'for sure rejectable'*.
- For inductive warrants (reasoning from examples to generalization) the assessment is likely to be '... *tolerable'* or '... *opposable'*, depending on the specific case whereas deductive warrants are likely to be assessed as '*for sure acceptable*'.

The proposed aggregation rules are given in table 1. They are expressed in terms of Dempster-Shaffer *belief* and *plausibility* functions. However, they can be easily expressed in terms of *confidence* and *decision* functions using equations (2) and (4).

Table 1: Aggregation Rules

A-argument rule Yager's modification of Dempster's rule of combination [12] (version for two arguments) $Bel(c) = Bel(a_1) \cdot Bel(a_2) + Bel(a_1) \cdot (Pl(a_2) - Bel(a_2)) + Bel(a_2) \cdot (Pl(a_1) - Bel(a_1))$ $Pl(c) = 1 - ((1 - Pl(a_1)) \cdot (1 - Pl(a_2)) + (1 - Pl(a_1)) \cdot (Pl(a_2) - Bel(a_2)) + (1 - Pl(a_2)) \cdot (Pl(a_1) - Bel(a_1)))$ Assessments coming from counter-arguments should be transformed before applying the rule of combination using the following equations: $Bel_{\arg ument}(a) = 1 - Pl_{counter-\arg ument}(a)$ $Pl_{\arg ument}(a) = 1 - Bel_{counter-\arg ument}(a)$
NSC-argument rule $Bel(c) = Bel(w) \cdot Bel(a_1) \cdot Bel(a_2) \cdot \ldots \cdot Bel(a_n)$ $Pl(c) = 1 - Bel(w) \cdot (1 - Pl(a_1) \cdot Pl(a_2) \cdot \ldots \cdot Pl(a_n))$
SC-argument rule $Bel(c) = Bel(w) \cdot Bel(a_1) \cdot Bel(a_2) \cdot \ldots \cdot Bel(a_n)$ $Pl(c) = 1$
C-argument rule

$$Bel(c) = Bel(w) \cdot \frac{k_1 Bel(a_1) + k_2 Bel(a_2) + ... + k_n Bel(a_n)}{k_1 + k_2 + ... + k_n}$$

$$Pl(c) = 1 - Bel(w) \cdot \left(1 - \frac{k_1 Pl(a_1) + k_2 Pl(a_2) + ... + k_n Pl(a_n)}{k_1 + k_2 + ... + k_n}\right)$$

Where:

- a_i is the i^th premise or argument strategy,
- k_i is a weight of the i^th premise defined by the argument proposer (it represents the coverage of the conclusion by the premise - see figure 2),
- *c* is a claim (conclusion),
- *w* is a warrant.

As experts are supposed to use the discrete (linguistic) scale shown in figure 2 and the aggregation rules are defined on the numeric scale [0,1]x[0,1], it is necessary to map the linguistic assessments to the numeric scale in order to apply the aggregation rules and then to map the results back, to present them to the users. Therefore, for each aggregation rule we define the *scaling function*:

$$s : X \rightarrow [0,1] \times [0,1] \qquad (5)$$

where *X* is a set of all the levels distinguished in the *Assessment scale*. This function had to be calibrated to make function $s^{-1} \circ f \circ s$, where $f : [0,1] \times [0,1] \rightarrow [0,1] \times [0,1]$ is the corresponding aggregation function presented in table 1, most closely matching the expert assessments of the whole argumentation.

5. EXPERIMENTAL CALIBRATION AND VALIDATION

The aim of the aggregation rules is to calculate the assessment of the conclusion in a way consistent with what would be deduced by an expert in such a situation. Therefore, appropriate experiments had to be performed to gather enough data to calibrate the scaling functions and to validate the proposed method.

A group of 31 students of the last year of a computer science university course took part in the experiment. The participants had already attended courses of logic and mathematics. They also attended a two-hour lecture about trust cases.

The students were divided into three groups. Each group was supposed to apply one of the aggregation rules: *A-rule, NSC-rule* and *C-rule*. *SC-argument* type has been dropped because of its similarity to *NSC*-argument type.

Each of the students was provided with 5 simple trust cases composed of a claim, an argument/arguments, a warrant/warrants and premises. The results of their assessments were collected in a pre-prepared questionnaire. The students were asked to assess the warrants and, in case of *C-argument* to assign weights to the premises. Then, assuming the pre-defined assessment of each premise (given in the questionnaire) the students were asked to give their assessment of the conclusion using the *Assessment Scale*. They were supposed to repeat this step for 10 different sets of initial assessments of the premises (chosen randomly) for each trust case. That makes 50 assessments of the conclusions in total issued by each student. To check for consistency, 10 randomly selected assessments were repeated for each student.

Some students were excluded from the experiment for formal reasons or because their assessments apparently were not reasonable (for instance, they declared high confidence in acceptance of a conclusion in a situation where the premises were with high confidence rejectable). Finally, 8 questionnaires related to *A-argument* type, 6 questionnaires related to *NSC*-argument type and 10 questionnaires related to *C*-argument type were used in the following analysis.

The data gathered were used to find the optimal scaling function for each type of aggregation rule.

In addition, the quality of each aggregation rule was assessed. To this end, consistency of the students' answers and accuracy of the estimation of their answers by the aggregation rules were calculated. Consistency was measured by calculating the root-mean-square value of the difference between the first and the repeated assessment of the same conclusion with the same values assigned to the premises.

Accuracy was measured by calculating the root-mean-square value of the difference between the student's assessment and the result of application of the aggregation rule. The above calculations were performed for both, *Confidence scale* and *Decision scale*. The results are presented in table 2. The numbers are normalized, which means that 1 represents the distance between two adjacent assessments on the linguistic scale.

The data show that the accuracy of the results obtained by application of the aggregation rules is similar to the consistency of participants' answers. This is the maximum of what could be achieved regarding the data set used to calibrate the aggregation rules. Further calibration requires more data which we plan to collect in the subsequent experiments.

Table 2: Results of the Experiment

Aggregation rule	**Consistency of students' assessments**		**Accuracy of assessments obtained by application of aggregation rules**	
	Confidence scale	***Decision scale***	***Confidence scale***	***Decision scale***
A-rule	1,03	0,64	1,06	0,80
NSC-rule	0,94	0,62	1,10	0,78
C-rule	0,84	0,87	0,90	0,66

6. CONCLUSION

This article introduced a method of gathering expert opinions about the validity of arguments and the value of the supporting evidence. The method provides means of assessing the compelling power of arguments contained in trust cases and communicating the results of such an assessment. The method is well suited to represent epistemic (subjective) uncertainty resulting from the lack of knowledge of the expert and to combine evidence of varying 'quality'. The method has been validated in some preliminary experiments. It has been implemented in the TCT system which supports full-scale application of Trust-IT [13] and is planned to be used in the assessment of real trust cases developed for e-health services in two 6th Framework EU research projects.

Acknowledgement

This work was partially supported by the project ANGEL - 'Advanced Networked embedded platform as a Gateway to Enhance quality of Life' (IST project 2005-IST-5-033506-STP) and by the project PIPS – 'Personalized Information Platform for health and life Services' (Contract No. 507019 IST2.3.1.11 e-Health) within the European Commission 6th Framework Programme.

References

[1] T. P. Kelly. "*Arguing Safety – A Systematic Approach to Managing Safety Cases*", PhD Thesis, University of York, UK, (1998).

[2] R. Bloomfield et al. "*Assurance Cases for Security*", A report from a Workshop on Assurance Cases for Security, Washington, USA, (2005).

[3] J. Gorski et al. "*Trust Case: Justifying Trust in IT Solution*", Elsevier, Reliability Engineering and System Safety, Volume 89, pp. 33-47, (2005).

[4] J. Gorski. "*Trust Case – a Case for Trustworthiness of IT Infrastructures*", Cyberspace Security and Defence: Research Issues, NATO ARW Series, Springer-Verlag, pp. 125-142, (2005).

[5] L. Cyra and J. Gorski. "*Supporting Compliance with Safety Standards by Trust Case Templates*", ESREL 2007, Norway, vol. 2, pp. 1367-1374, (2007).

[6] L. Cyra and J. Gorski. "*Standard Compliance Framework for Effective Requirements Communication*", Polish Journal of Environmental Studies, Volume 16 no. 5B, pp. 312-316, (2007).

[7] L. Cyra and J. Gorski. "*Extending GQM by Argument Structures",* 2nd IFIP Central and East European Conference on Software Engineering Techniques CEE-SET, (2007).

[8] L. Strigini. "*Formalism and Judgement in Assurance Cases*", Workshop on Assurance Cases: Best Practices, Possible Obstacles, and Future Opportunities, Proc. of DSN 2004, Italy, (2004).

[9] J. Gorski. "*Trust-IT – a Framework for Trust Cases"*, Workshop on Assurance Cases for Security: The Metrics Challenge, Proc. of DSN 2007, UK, (2007).

[10] J. Gorski and M. Zagorski. "*Reasoning about trust in IT infrastructures*", ESREL 2005, (2005).

[11] G. Shafer. "*Mathematical Theory of Evidence*", Princetown University Press, (1976).

[12] K. Sentez and S. Ferson. "*Combination of Evidence in Dempster-Shafer Theory*", SANDIA National Laboratories, (2002).

[13] Information Assurance Group, "*TCT User Manual*", Gdansk University of Technology, http://kio.eti.pg.gda.pl/trust_case/download/TCTEditor_Users_Manual.pdf, (2007).