Postprint of: Cyra Ł., Górski J., Support for argument structures review and assessment, Reliability Engineering & System Safety, Vol. 96, Iss. 1 (2011), pp. 26-37, DOI: 10.1016/j.ress.2010.06.027

Support for Argument Structures Review and Assessment

Lukasz Cyra[1], Janusz Górski[2*]

[1] Adelard LLP
College Building, 10 Northampton Square, London EC1V 0HB, UK
lc@adelard.com

[2] Gdansk University of Technology, Department of Software Engineering
Narutowicza 11/12, 80-952 Gdansk, Poland
jango@pg.gda.pl

Abstract. Argument structures are commonly used to develop and present cases for safety, security and other properties of systems. Such structures tend to grow excessively, which causes problems with their review and assessment. Two issues are of particular interest: (1) systematic and explicit assessment of the compelling power of an argument, and (2) communication of the result of such an assessment to relevant recipients. The paper presents a solution to these problems. The method of *Visual Assessment of Arguments (VAA)*, being this solution, is based on the Dempster-Shafer theory of evidence applied to assessment of the strength of arguments, and a visual mechanism of issuing and presenting assessments, supported by the so called opinion triangle. In the paper we explain theoretical grounding for the method and provide guidance on its application. The results of some validation experiments are also presented.

Keywords: Argument structures, Argument assessment, Dempster-Shafer model, VAA, Trust Case.

1 Introduction

Arguments are commonly used in 'cases' (safety cases [18, 22, 23], assurance cases [3], trust cases [13, 14], conformity cases [7, 8], etc.) to justify various qualities of objects (like safety, security, privacy, conformity with standards and so on). Recently, there is a growing interest in these subjects, which leads to the development of relevant methodologies and finding new application areas for argument structures [9, 10].

The idea behind the development of argument structures is to make expert judgment explicit in order to redirect the dependence on judgment to issues on which we can trust this judgment [29]. Presenting explicitly an argument together with the evidence that supports it makes it is possible to analyse the argument by third parties

* Corresponding author. Tel: +48583471909, fax: +48583472727, E-mail: jango@pg.gda.pl

and take a position on it. However, argument structures tend to grow excessively, becoming too complex to be analyzed by non-experts. Therefore, appropriate methods of assessment of argument structures are required. Two objectives are of particular interest: (1) assessment of the compelling power of an argument, and (2) communication of the result of such an assessment to relevant recipients.

To address these issues the paper introduces a new method of appraising arguments, which is called the Visual Assessment of Arguments (VAA). VAA is general enough to be applied in any situation, in which arguments are presented explicitly together with supporting evidence. It provides for issuing assessments of evidence referred to in the argument and of the inference rules applied in it. The method proposes linguistic appraisal scales, which are mapped onto the Josang's opinion triangle [20, 21] so that assessments can be issued also visually. The assessments (of the evidence and of the inference rules) are represented in terms of Dempster-Shafer belief functions [27, 28] and aggregated leading to the assessment of the whole argument.

In this paper VAA is presented in relation to the Trust-IT methodology [12, 13, 14], and all the examples of its application refer to trust cases. Some of the figures included in the paper were created using the TCT (Trust Case Toolbox) software tool [19], which fully implements VAA. The paper also includes examples of the application of the method, which have been borrowed from a trust case developed for a real system [26].

VAA and its implementation in TCT have already been subjected to excessive validation. We report on the results of experiments carried out to validate and calibrate the argument appraisal method. Furthermore, the method has already been applied with respect to several real-world cases (focusing on safety, security and privacy) in two EU funded research projects [1, 26].

2 Related Work

The problem of correctness of and confidence in argument structures have been already addressed by some authors (e.g. [4, 31]). However, the methods which effectively support argument structures review and assessment are continually sought after.

The present research in this area takes three different routs, which have slightly different objectives and scopes.

Structural correctness (rules, checklists). This group of methods is based on checklists and rules which support auditors in argument structures review. These tools may help in discovering flaws in logic of a proof (e.g. as in P. Mayo's systematic approach to safety case review [25]). They can also capture expert knowledge identifying common patterns of argumentation and describing the most common fallacies in their implementation (e.g. following W. Greenwell, J. Knight, C. Holloway and J. Pease's taxonomy of fallacies in safety system arguments [17]). Such methods are effective in identifying certain types of problems. They concentrate on the inference applied in arguments but, in particular, do not address strength and validity of evidence.

Quality Models. This group of methods similarly to the previous one aims at identifying potential problems in argument structures. One of such methods is being developed by A. Zechner and M. Huhn [33]. Here, information from argument structures is used to build quality models, which support the identification of inconsistencies and tacit trade-offs between the activities referred to from the argument in relation to its claims. This way faulty or missing links between claims and premises can be found and the argument can be improved. These methods, however, similarly to the previous ones do not address strength and validity of evidence.

Quantitative approaches. This group of methods tries to apply mathematical formalism to capture the uncertainty, resulting from the lack of evidence or the argument fallacies, related to stated claims. Two main ways of approaching the problem can be identified:

- **Bayesian Belief Networks.** BBNs are applied to deduce the confidence in a goal from credibility of its backing arguments. Examples of such approach are W. Wu and T. Kelly [32] and B. Littlewood and D. Wright [24]. They require much expertise in BBNs from the users and a lot of input prior to generating any conclusions.
- **Approaches based on the Dempster-Shafer theory of evidence** [27, 28]**.** These approaches capture information (assessments) about validity and strength of evidence using simple scales and then aggregate this information to obtain conclusions about the whole argument. The VAA method presented in this paper belongs to this group.

The first two groups are purely qualitative and can be considered, to a great extent, complementary to the quantitative approaches (based on BBN and the Dempster-Shafer theory).

We believe that methods based on the Dempster-Shafer theory are much more promising than the BBN based approaches because their application is much easier.

As for the Dempster-Shafer based approaches only two methods exist: VAA and [16]. VAA significantly extends and improves the argument appraisal method proposed in [16]. The extensions and improvements are as follows:

- Coverage of the whole range of inference rules (in [16] only two situations were supported: (1) aggregation of assessments from facts to a conclusion, and (2) aggregation of assessments from sub-claims to a conclusion, which resulted in problems with adapting the method to different types of inference used in arguments)
- Uniform mechanism of aggregation of assessments, the same for all kinds of argument structure nodes (in [16] different representations of the same arguments lead to different assessments after aggregation)
- Coverage of the whole range of node types
- More clear assessment criteria, focusing on the information contained in the node, uniform for different locations of nodes in the argument structure
- Application of linguistic assessment scales, which are expressed in terms understandable by the user (in [16] numeric scales are used, the interpretation of which is not defined)
- Experimental calibration and validation of the aggregation mechanism

2 Representing Arguments

The proposed approaches to argument representation in 'cases' [2, 13, 22] are influenced by Toulmin's argument model [30]. In our approach (the Trust-IT methodology [12, 13, 14]) we adopt this model in a fairly straightforward way as shown in Fig. 1.

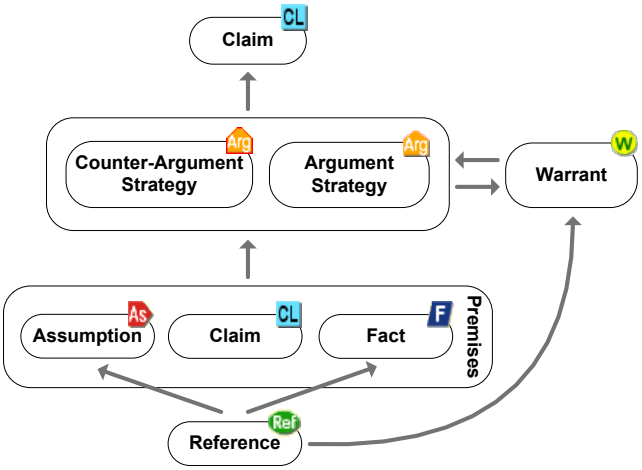

Fig. 1 Trust-IT argument model.

The presented structure includes a conclusion to be justified, represented as a *claim* (denoted CL). The claim is supported by an *argument strategy* (denoted Arg), which contains a basic idea how to support the conclusion. In the case of *counter-argument strategies* (denoted Arg) it includes the idea of rebuttal of the claim. The argument strategy is related to a *warrant* (denoted W), which justifies the inference from *premises* to the conclusion. This justification may require additional, more specific arguments, which is shown by the arrow leading from the argument to the warrant node.

A premise can be of three different types: it can be 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. Evidence is provided in external (to the trust case) documents, which are pointed at by nodes of type *reference* (denoted Ref). In the case of an assumption, the referenced document can contain explanation of the context in which the assumption is made.

As claims and warrants can be demonstrated using other (sub-)claims the argument structure can grow recursively.

The icons for different nodes, which are shown in Fig.1 and implemented in the supporting tool (TCT [19]), will be used in the subsequent examples presented in this paper.

An example argument following the model introduced in Fig. 1 is presented in Fig. 2. The example refers to the PIPS system, which delivers to its users health and lifestyle related personalized services [26].

Top claim *C0*, postulates validity of information supplied to PIPS. It is demonstrated by considering different channels through which information related to a patient's state is supplied to the system (argument strategy *A0* and warrant *W0*). This leads to four premises which are used by the argument. Three of them: *C1*, *C2* and *C3* are claims and are supported by more detailed arguments, and the fourth one A1 is an assumption and is not analyzed further.

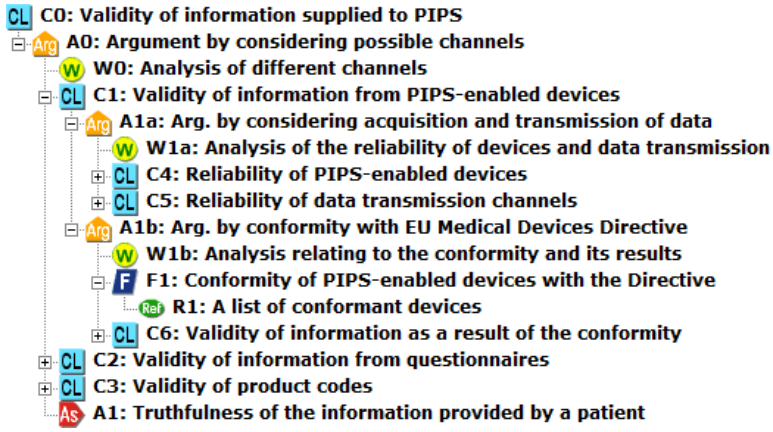

Fig. 2 Example argument related to trustworthiness of the PIPS system.

Claim *C1* is supported by two arguments: *A1a* and *A1b*.

In warrant *W1a*, it is argued that the information from the PIPS-enabled devices is reliable if the devices themselves are reliable (claim *C4*) and the channels of data transmission are reliable (claim *C5*).

In warrant *W1b*, it is argued that the information from the PIPS-enabled devices is reliable if the PIPS-enabled devices are conformant with the European Medical Devices Directive (fact *F1*) and the validity of information from the devices is implied by the conformity with the directive (claim *C6*).

The question related to the argument presented in Fig. 2 is: '*How well does it support the topmost claim*?' VAA proposed in the following sections helps answer this question by calculating the support given to a claim based on the expert assessments of the basic elements of the argument: assumptions, facts and warrants.

3 Properties of Arguments

Before presenting the VAA method in detail, we first introduce the assumptions which have influence on the selection of aggregation rules applied in our method:

Assumption 1. From the appraisal mechanism viewpoint, conclusions and premises of arguments are considered to be sentences. They are assessed using the same criteria which reflect the acceptability of a given sentence

by the assessor and the confidence the assessor has in this judgement. Therefore, the appraisal mechanism treats claims, facts and assumptions the same way.

Assumption 2. As arguments have a tree-like structure, the assessment starts from the most detailed premises (facts and assumptions being the leaves of the tree) and then recursively traverse the tree inferring the assessments of the conclusions (claims).

Assumption 3. A claim can be supported by diverse argument strategies, including counter-arguments. Therefore, the appraisal mechanism provides means for aggregation of the assessments resulting from different arguments (dealing with possible contradictions).

Assumption 4. The support given to a conclusion differs depending on the inference rule proposed in the related warrant. Therefore, the appraisal mechanism aggregates assessments in a warrant-dependent way.

Assumption 4 requires more detailed explanation. We distinguish two main types of inference rules occurring in argument structures:

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. Examples are warrants *W1a* and *W1b* in Fig. 2.

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 (possibly with less confidence). Example is warrant *W0* in Fig. 2.

It may be observed that many 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 rejection of the conclusion (i.e. in the situation where one of the premises is false the conclusion is false). This type of an argument is called *NSC-argument (Necessary and Sufficient Condition list argument)*. (For instance, warrant *W1a* Fig. 2.)

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 (i.e. in the situation where one of the premises is false nothing can be inferred about the conclusion). This type of an argument is called *SC-argument (Sufficient Condition list argument)*. (For instance, warrant *W1b* in Fig. 2.)

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

Arguments of Type 2 can be considered further. To this end, let us consider the example presented in Fig. 3.

In the example, each of the claims: *C1*, *C2* and *C3* states that a person is healthy.

Claim *C1* is argued based on 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 (these references are not shown in Fig. 3).

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 assessed without performing any sophisticated examinations. The latter refers to aggregated results of different examinations performed by different specialists, which altogether provide detailed presentation of the patient's health state.

Claim *C3* (similarly to *C1*) demonstrates health by reference to the 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*).

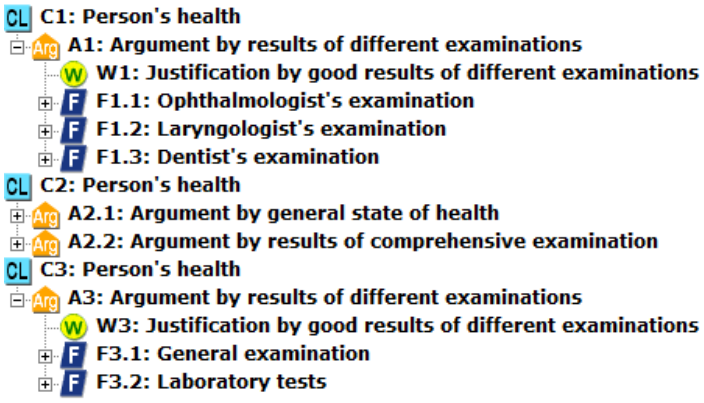

Fig. 3 Examples of arguments of Type 2

Differences between the arguments supporting claims *C1*, *C2* and *C3* in Fig. 3 are illustrated in Fig. 4.

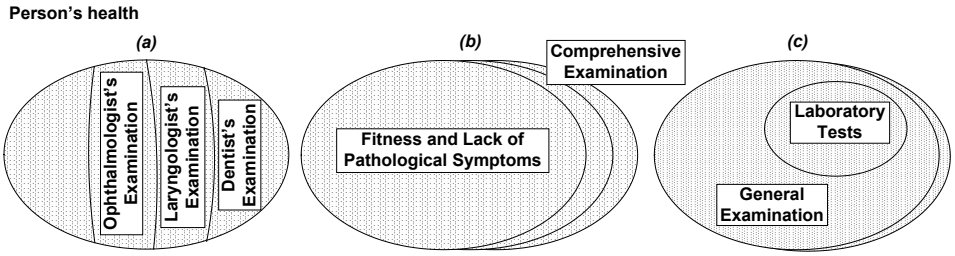

Fig. 4 Sub-categories of the argument of Type 2

As shown in Fig. 4, arguments of Type 2 can be divided into 3 sub-categories:

Type 2.1: Each of the premises 'covers' part of the conclusion, as represented in Fig. 4(a). This models argument *A1* for claim *C1* in Fig. 3. 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: The premises are used in several independent arguments and support the whole conclusion, as represented in Fig. 4(b). This models arguments *A2.1* and *A2.2*

for claim *C2* in Fig. 3. 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), as represented in Fig. 4 (c). This models argument *A3* of Fig. 3. Laboratory tests assess some aspects of a patient's health and overlap with the results obtained from the general examination. 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. Therefore, a pragmatic solution has been adopted that if the overlapping is considered insignificant, an argument of Type 2.3 is treated as a C-argument and, if the overlapping is considered significant, the argument is treated as an A-argument.

4 Argument Appraisal Procedure

Now we are ready to explain the argument appraisal procedure of VAA. It comprises the following steps.

Step 1 – appraisal of warrants and premises. This step is split into two sub-steps:

1.1 Assess basic warrants (the warrants which are not justified by explicit arguments) occurring in the argument. This assessment is based on the assessment of the evidence linked to the warrant (if any) but also refers to the common knowledge and logic bases for the inference.

1.2 Assess the facts and assumptions occurring in the argument. This appraisal is mostly based on the assessment of the evidence linked to the premises by the reference nodes.

Referring to the example shown in Fig. 2, the appraisal of the *'W0: Analysis of different channels'* warrant would take into account if the validity of information received from the devices, questionnaires (with the additional assumption that patients are not cheating intentionally) and by reading product codes is sufficient to conclude the validity of the information supplied to the system.

The appraisal of the premises would assess the acceptability of the assumption that patients are not cheating intentionally (note that this is context dependent and the result would depend on the knowledge about the system and its environment). The appraisal of the *'F1: Conformity of the PIPS-enabled devices with European Medical Devices Directive'* fact would take into account the evidence linked to this fact by the corresponding reference node.

Implementation of Step 1 requires that an appropriate assessment scale to express the appraisals of warrants, facts and assumptions is available.

Step 2 – aggregation of the partial appraisals. This is performed in the following two sub-steps:

2.1 For each claim, whose all premises and the related warrant possess an appraisal; aggregate the appraisals of premises and the warrant to obtain the appraisal of the claim.

2.2 Repeat step 2.1 until the top claim is reached.

Referring to the example from Fig. 2, this step would result in the appraisal of the top claim taking as an input the appraisals of warrants, facts and assumptions occurring in the argumentation and recursively applying the aggregation rules.

Implementation of Step 2 requires that appropriate aggregation rules are defined, covering all relevant types of warrants occurring in arguments.

5 Assessment Scale

To support experts during the appraisal process two linguistic scales have been introduced, the *Decision scale* and *Confidence scale*. The former provides for expressing the attitude towards acceptance or rejection of the assessed element. It comprises four decision values: *'acceptable'*, *'tolerable'*, *'opposable'* and *'rejectable'*. The latter provides for expressing the confidence in this decision. It distinguishes six levels of confidence: *'for sure'*, *'with very high confidence'*, *'with high confidence'*, *'with low confidence'*, *'with very low confidence'* and *'lack of confidence'*.

The scales can be combined together which results in twenty-four values of the Assessment scale, as shown in Fig. 5. The elements of the scale, which are represented as small circles, have intuitively understandable linguistic values. For instance, the element represented in Fig. 5 as a hollow circle reads: '*with very low confidence tolerable'*.

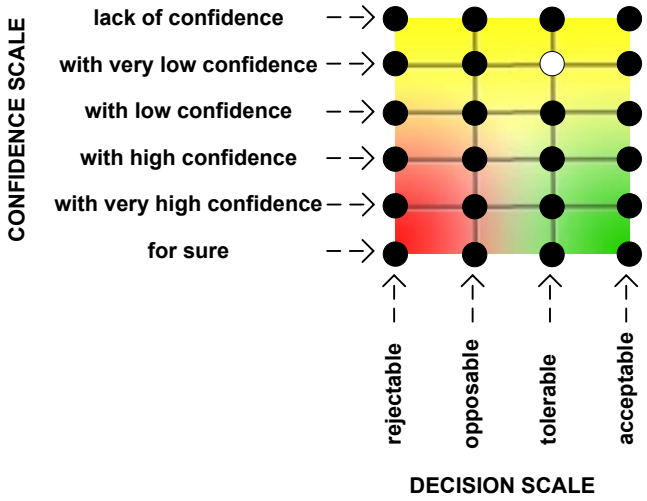

Fig. 5 Combined Confidence and Decision scales

The semantics of the scales can be formalized using Dempster-Shafer's belief and plausibility functions [27, 28].

If *s* is a statement, then

$Bel(s) \in [0,1]$ is the *belief* function representing the amount of belief that directly supports *s*,

$Pl(s) \in [0,1]$ is the *plausibility* function representing the upper bound on the belief in *s* that can be gained by adding new evidence.

Referring to the above functions we formally define *confidence* in the following way:

$$Conf(s) = Bel(s) + 1 - Pl(s)$$
$$Conf(s) \in [0,1] \tag{1}$$

The linguistic values from the Confidence scale are mapped into the [0,1] interval in such a way that *'lack of confidence'* = 0, *'for sure'* = 1 and the other linguistic values are distributed evenly in the interval. The opposite transformation can be performed by choosing the closest linguistic value which corresponds to a particular numerical value.

The 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 the decision concerning *s* can be formally represented as:

$$Dec(s) = \begin{cases} \dfrac{Bel(s)}{Bel(s) + 1 - Pl(s)} & Bel(s) + 1 - Pl(s) \neq 0 \\ 1 & Bel(s) + 1 - Pl(s) = 0 \end{cases} \tag{2}$$

$$Dec(s) \in [0,1]$$

The linguistic values from the Decision scale are mapped into the [0,1] interval in such a way that '*rejectable'* = 0, *'acceptable'* = 1 and the other linguistic values are distributed evenly in the interval. The opposite transformation can be performed by choosing the closest linguistic value which corresponds to a particular numerical value.

The transformation from the confidence and decision functions to the Dempster-Shafer belief and plausibility functions is defined by the following equations:

$$Bel(s) = Conf(s) \cdot Dec(s)$$
$$Pl(s) = 1 - Conf(s) \cdot (1 - Dec(s)) \tag{3}$$

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

The difference between stating that something is acceptable or rejectable is significant if enough (or at least some) evidence supporting this decision is available. For instance, *'for sure acceptable'* or *'with very high confidence rejectable'* needs to be based on the evaluation of the available evidence.

In the case of *'lack of confidence'* the situation is different, however. Lack of confidence refers to the situation, where there is no evidence available (or the

evidence is irrelevant) and therefore it does not matter which value is chosen from the Decision scale. In other words, there is no reason to distinguish between '*with lack of confidence acceptable*' and '*with lack of confidence rejectable*' as both assessments express complete uncertainty about the corresponding decision. Therefore, in such situation all four elements from the Decision scale are treated as equivalent and the assessment is simply called *'lack of confidence'*. This is illustrated in Fig. 6, where all bullets related to '*lack of confidence*' are merged into one. The result is a triangle which we call the *Assessment Triangle*. It corresponds to the Josang's opinion triangle [20, 21] where *'lack of confidence'* is mapped onto uncertainty and the other vertices represent the total disbelief (equivalent to the '*for sure rejectable'*) and total belief (equivalent to the '*for sure acceptable'*).

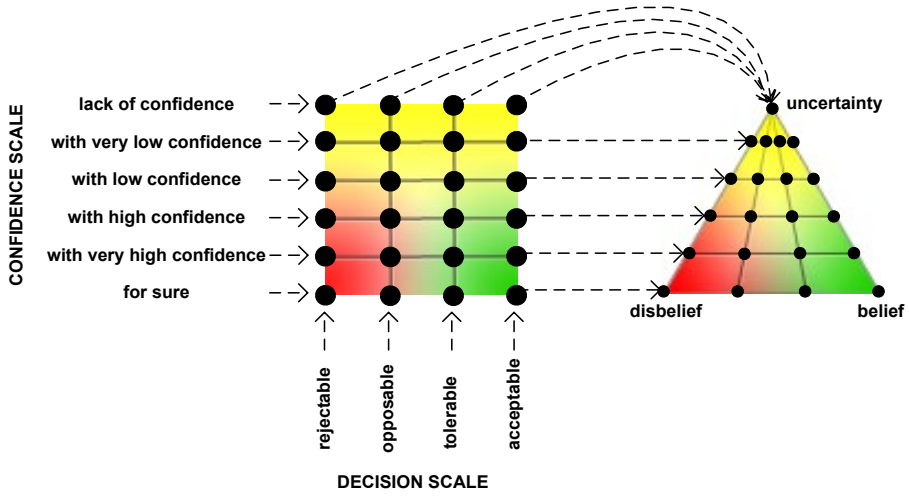

Fig. 6 The Assessment Triangle

6 Procedure of issuing assessments

The Assessment Triangle is applied to express opinions and the level of confidence in these opinions in relation to the elements (assumptions, facts, warrants) of an argument structure.

Assessment of a single element proceeds as follows:

Step 1 - If no evidence for or against the statement representing the node is available the *'lack of confidence'* assessment is issued and the procedure terminates.

Step 2 - In the opposite case, the ratio between the evidence supporting the acceptance and rejection of the statement is assessed and an appropriate value from the Decision scale is chosen.

Step 3 - Then, it is assessed how much evidence could be additionally provided to become sure about the decision taken in step 2. This amount of missing evidence drives the selection from the Confidence scale.

Step 4 - The final assessment from the Assessment Triangle is obtained by combination of the two assessments from Step 2 and Step 3.

Fig. 7 presents the user interface for issuing assessments of the TCT tool [19]. The user can drag a small hollow marker over the Assessment Triangle shown on the left hand side. Then, the linguistic values corresponding to the current position of the marker are displayed below the *Confidence level* and *Decision* sliders. It is also possible to directly choose an appropriate linguistic assessment in each window. Additionally, the current levels of belief, disbelief and uncertainty are displayed as horizontal bars just above the opinion triangle.

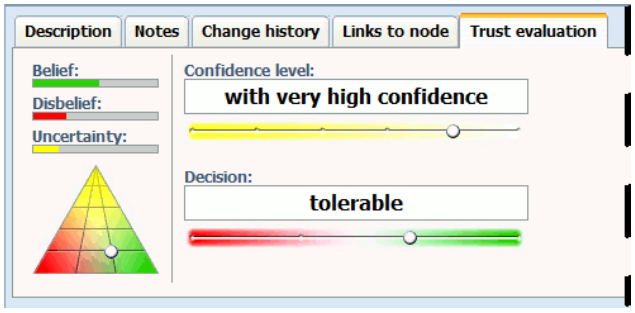

Fig. 7 User interface for issuing assessments of TCT

7 Examples

As an example let us consider fact *F1* from Fig. 2 stating that PIPS-enabled devices are conformant with the European Medical Devices Directive.

Let us assume that:

(1) There is some evidence E related to the fact.

(2) E supports *F1*, which results in choosing the *'acceptable'* value from the Decision scale.

(3) evidence demonstrates well fulfilment of the requirements of the directive, however, the formal certification process has not been performed yet, which leads to the *'with very high confidence'* assessment.

Consequently, the final appraisal of *F1* is: *'with very high confidence acceptable'*.

Let us now consider *F1* in a different situation:

(1) There is some evidence E related to the fact.

(2) E demonstrates that most of the requirements of the directive are met apart from one significant requirement – this results in the *'opposable'* assessment.

(3) E is substantial, however, not complete, which gives the *'with high confidence'* assessment.

Consequently, the final appraisal of *F1* is: *'with high confidence opposable'*.

As another example let us take warrant *W0* (*'Analysis of different channels'*) from Fig. 2. The warrant identifies different types of channels providing information to the system and explains that if they provide valid information, the information available to the system is also valid.

Let us assume that:

(1) An inventory of the channels exists.

(2) It identifies the four major types of channels represented in the argument; in addition, some other channels (considered less important), which were not taken into account in the argument – this leads to the *'tolerable'* assessment of the warrant.

(3) The inventory results from a formalized procedure of system review – this leads to the *'for sure'* assessment.

Consequently, the final appraisal of *W0* is: *'for sure tolerable'*.

As yet another example let us take assumption *A1* (*'Truthfulness of the information provided by a patient'*) from Fig. 2, which states that patients will not intentionally input false data into the system.

Let us assume that:

(1) Information about the range of possible patient interactions and the scope of data input into the system is available.

(2) The assessor tends to accept the assumption, however, considers also some situations where it not necessarily holds – the resulting assessment is: *'tolerable'*.

(3) The assessor has no doubts that he/she sees the whole scope of relevant situations – this leads to the *'for sure'* assessment.

Consequently, the final appraisal of *A1* is: *'for sure tolerable'*.

8 Aggregation Rules

Aggregation rules define how the assessments of the premises and the assessment of the warrant are used to calculate the appraisal of the conclusion.

The following four rules of aggregation are distinguished. They are related to the argument types identified in section 3:

- *C-argument Rule* - to calculate the assessment of the conclusion for C-arguments,
- *NSC-argument Rule* - to calculate the assessment of the conclusion for NSC-arguments,
- *SC-argument Rule* - to calculate the assessment of the conclusion for SC-arguments,
- *A-argument Rule* - to calculate the assessment of the conclusion for A-arguments.

It is assumed that each warrant occurring in the argumentation has its type explicitly identified and the corresponding aggregation rule assigned, which is done by the developer of the argument structure. Additionally, for C-arguments it is assumed that the argument developer defines weights assigned to the premises, which indicate the relative support given by a premise to the conclusion (see Fig. 4(a)).

The proposed aggregation rules are expressed in terms of Dempster-Shafer's belief and plausibility functions. They can be easily expressed in terms of confidence and decision functions using the transformation equations (1), (2) and (3).

In the sequel we assume that:

- c is a claim (conclusion),
- w is the warrant of c (in the case c has only one argument strategy),
- n is the number of premises/argument strategies of c,
- a_i (where $i \in \{1, \ldots, n\}$) is the i[th] premise,
- ar_i (where $i \in \{1, \ldots, n\}$) is justification contained in the i[th] argument strategy,
- k_i (where $i \in \{1, \ldots, n\}$) is the weight assigned to the i[th] premise.

To understand the equations presented below it should be remembered that if s is a statement, then

- $1 - Pl(s)$ represents the total disbelief in s, and
- $Pl(s) - Bel(s)$ represents the total uncertainty related to s.

8.1 A-Argument Rule

A-argument relates to a situation where more than one independent justification (argumentation branch) of the common conclusion is provided. In A-arguments, confidence in assessments coming from different argumentation branches is reinforced if the assessments agree, or it is decreased if they contradict each other.

For the A-argument, Yager's modification of Dempster's rule of combination [27] is applied. Below the version of this rule for two argument strategies is presented:

$$\begin{aligned} Bel(c) &= Bel(ar_1) \cdot Bel(ar_2) + Bel(ar_1) \cdot [Pl(ar_2) - Bel(ar_2)] \\ &\quad + Bel(ar_2) \cdot [Pl(ar_1) - Bel(ar_1)] \\ Pl(c) &= 1 - \{[1 - Pl(ar_1)] \cdot [1 - Pl(ar_2)] + [1 - Pl(ar_1)] \cdot [Pl(ar_2) - Bel(ar_2)] \\ &\quad + [1 - Pl(ar_2)] \cdot [Pl(ar_1) - Bel(ar_1)]\} \end{aligned} \tag{4}$$

In the above equations the belief in the conclusion is a sum of three components each of which represents the belief built on: (1) the belief in both of the arguments; (2) the belief in the first argument and uncertainty in the second one; and (3) the belief in the second argument and uncertainty in the first one.

Plausibility of the conclusion is calculated as 1 minus the total disbelief in the conclusion (obtained by analogical calculations as for obtaining the belief in it).

When more than two argument strategies exist, appropriate modifications of these equations should be applied.

Assessments coming from counter-arguments are first being transformed, before applying rule (4), using the following equations:

$$\begin{aligned} Bel_{\text{arg}ument}(ar_i) &= 1 - Pl_{counter-\text{arg}ument}(ar_i) \\ Pl_{\text{arg}ument}(ar_i) &= 1 - Bel_{counter-\text{arg}ument}(ar_i) \end{aligned} \tag{5}$$

where $i \in \{1, \ldots, n\}$.

To better understand implications of these equations let us illustrate the situation with an example shown in Fig. 8. The assessments coming from argument strategies are shown on the right hand side in italic. Both arguments support the conclusion providing high confidence. In this case, the resultant assessment of the conclusion is *'with very high confidence acceptable'* (in bold italic).

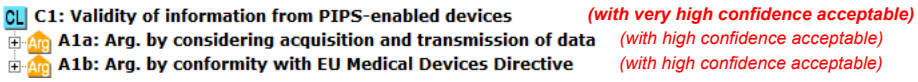

Fig. 8 Assessment of the conclusion of A-argument

In the case the arguments contradicted each other, the effect would be opposite. If one of the arguments in Fig. 8 supported rejection of the conclusion and another recommend acceptance, both with the same level of confidence, there would be no confidence in the conclusion and the *'lack of confidence'* assessment would result.

8.2 C-Argument Rule

C-argument relates to a situation where the premises provide complementary support for the conclusion. In such a case not only the assessments of the premises and the warrant but also the weights associated with each premise are taken into account. The final assessment of the conclusion is a sort of weighed mean value of the contribution of all the premises.

For the C-argument, the following aggregation functions are proposed:

$$Bel(c) = Bel(w) \cdot \frac{k_1 Bel(a_1) + k_2 Bel(a_2) + \ldots + k_n Bel(a_n)}{k_1 + k_2 + \ldots + k_n}$$
$$Pl(c) = 1 - Bel(w) \cdot \left(1 - \frac{k_1 Pl(a_1) + k_2 Pl(a_2) + \ldots + k_n Pl(a_n)}{k_1 + k_2 + \ldots + k_n}\right) \quad (6)$$

In equations (6) the belief in the conclusion is a weighed sum of the beliefs in each of the premises multiplied by (i.e. diminished to) the belief in the warrant.

Plausibility of the conclusion is calculated similarly i.e. a weighed sum is computed. Then a total disbelief is computed (1 – the weighed sum), which is multiplied by the belief in the warrant (to appropriately diminish the level of confidence), and again subtracted from 1 to obtain plausibility.

As an example, let us consider the C-argument shown in Fig. 9. The assessments of the warrant and premises together with the associated weights are shown on the right hand side in italic. The resulting assessment of the conclusion is *'with high confidence acceptable'*. Note that despite the fact that one of the premises is *'tolerable'* the conclusion is *'acceptable'*. This results from the fact that the other premises ranked *'acceptable'* outweighed in this case. Additionally, it can be seen that the confidence in the conclusion is slightly lower than it could be expected while looking at the assessments of the premises. This results from the fact that there were some doubts concerning the strength of the inference rule, reflected in the assessment of the warrant.

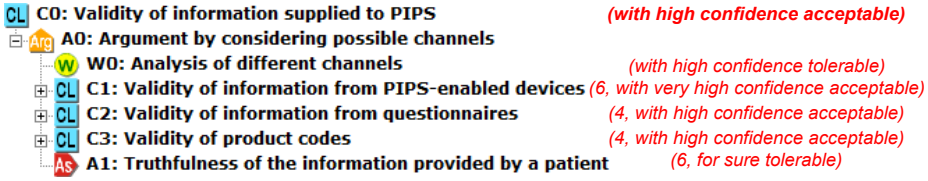

Fig. 9 Assessment of the conclusion of C-argument

Let us consider the example from Fig. 9 again but with the assessment of assumption *A1* modified to *'lack of confidence'*. In such a case the assessment of the conclusion would be *'with low confidence acceptable'*, which results from the fact that the other premises are highly assessed and there is relatively high assessment of the warrant. Nevertheless, the assessment of the conclusion would be lower than in the example shown in Fig. 9.

8.3 NSC-Argument Rule

In NSC-arguments, negative assessments are strongly reinforced. In such arguments the acceptance of all premises leads to the acceptance of the conclusion, whereas rejection of a single premise leads to the rejection of the conclusion.

For the NSC-argument the following family of functions is proposed:

$$\begin{aligned} Bel(c) &= Bel(w)\cdot Bel(a_1)\cdot Bel(a_2)\cdot\ldots\cdot Bel(a_n) \\ Pl(c) &= 1-Bel(w)\cdot[1-Pl(a_1)\cdot Pl(a_2)\cdot\ldots\cdot Pl(a_n)] \end{aligned} \tag{7}$$

In the equations (7) the belief in the conclusion is a product of the beliefs in each of the premises and the warrant. This way, each of the elements of the product if low, diminishes the resultant belief significantly.

Plausibility is calculated as 1 minus the whole disbelief coming from the premises multiplied by (i.e. diminished to) the belief in the warrant. The whole disbelief coming from the premises, on the other hand, is calculated as 1 minus the belief that each of the premises is trustworthy or uncertain.

An example of such an argument is shown in Fig. 10. Each premise is a necessary condition for the conclusion. Therefore, if even one of them is rejected, the conclusion cannot be accepted.

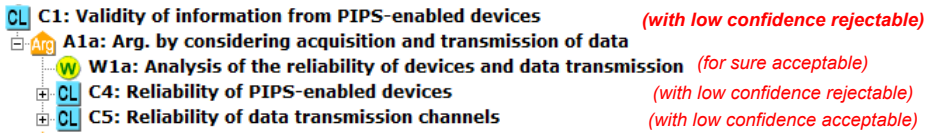

Fig. 10 Assessment of the conclusion of NSC-argument

Consequently, low assessments of the premises lead to a rapid drop in the assessment of the conclusion. If the second of the premises in the example were also

assessed as *'with low confidence rejectable'* the conclusion would be *'with high confidence rejectable'*.

8.4 SC-Argument Rule

In SC-arguments, acceptance of the premises leads to the acceptance of the conclusion, as for NSC-arguments. However, rejection of a single premise leads to the rejection of the whole inference, i.e. to complete lack of confidence.

For the SC-argument the following family of functions is proposed:

$$\begin{aligned} Bel(c) &= Bel(w) \cdot Bel(a_1) \cdot Bel(a_2) \cdot \ldots \cdot Bel(a_n) \\ Pl(c) &= 1 \end{aligned} \tag{8}$$

In the equations (8), similarly to the NSC-rule, the belief in the conclusion is a product of the beliefs into each of the premises and the warrant. This way, each of the elements of the product, if low, diminishes the resultant belief significantly.

Plausibility, however, is always set to 1. This means that this type of inference cannot lead to the rejection of the conclusion.

An example of such an argument is presented in Fig. 11. The lack of conformity with the EU Medical Devices Directive does not result in invalid information received from the devices. The only reasonable conclusion is that in such a case we do not know anything new concerning the validity of this information. Therefore, despite the fact that one of the premises is only *'tolerable'* the conclusion is *'acceptable'*, however, the level of confidence in this assessment is much lower than in each of the premises and in the warrant.

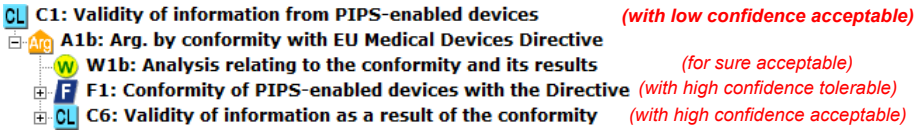

Fig. 11 Assessment of the conclusion of SC-argument

9 Experimental Evaluation

The Confidence and Decision scales support assessors by offering them a (not too large) set of intuitively understood linguistic values. However, in order to perform calculations defined in the aggregation rules, it is necessary to map the linguistic values onto the $[0,1]$ interval. This mapping can have a significant impact on the computed results.

There was no evidence that the (most obvious) even distribution of each of the linguistic scales over the $[0,1]$ interval is the most proper one. Therefore we decided to find the mapping experimentally. The aim of the experiment was to calibrate the

aggregation rules by relating their results to the expert assessments of the conclusions of a selected set of arguments. The experiment was conducted as follows.

For each aggregation rule, a scaling function *s* was defined, where

$$s:[0,1]\times[0,1]\rightarrow[0,1]\times[0,1] \tag{9}$$

takes as input a pair $(Conf(a), Dec(a))$, where *a* is a given statement and $Conf(a)$ and $Dec(a)$ are the numeric values corresponding to the chosen linguistic values from the Confidence and Decision scales assuming the even distribution of the linguistic values over the interval [0,1]. The output of *s* delivers a pair of numeric values which represent the chosen linguistic values and not necessarily follow the even distribution principle (which means for instance, that the distance on the numeric scale between *'for sure'* and '*with very high confidence'* is not necessarily the same as the distance between *'with very low confidence'* and '*lack of confidence'*).

A separate scaling function for each aggregation rule was defined in order to provide means of calibration of the aggregation functions apart from calibration of the scales only.

For each aggregation rule $aggr(as_1, as_2, \ldots, as_n)$ (where as_i is a pair of assessments of the aggregation rule's i[th] parameter expressed as numerical values in the even distribution, *n* is the number of parameters), the corresponding function *S* was calibrated to make the function $s^{-1}(aggr(s(as_1), s(as_2), \ldots, s(as_n)))$, most closely matching the expert assessments of the arguments corresponding to $aggr()$.

The general form of function $s()$ was assumed as follows:

$$s(x,y) = < s_{Conf}(x), s_{Dec}(y) >$$

$$s_{Conf}(x) = \begin{cases} c_1 + 5(1-c_1)(x-4/5) & 4/5 \le x \le 1 \\ c_2 + 5(c_1-c_2)(x-3/5) & 3/5 \le x < 4/5 \\ c_3 + 5(c_2-c_3)(x-2/5) & 2/5 \le x < 3/5 \\ c_4 + 5(c_3-c_4)(x-1/5) & 1/5 \le x < 2/5 \\ 5c_4 x & 0 \le x < 1/5 \end{cases} \tag{10}$$

$$s_{Dec}(y) = \begin{cases} d_1 + 3(1-d_1)(y-2/3) & 2/3 \le y \le 1 \\ d_2 + 3(d_1-d_2)(y-1/3) & 1/3 \le y < 2/3 \\ 3d_2 y & 0 \le y < 1/3 \end{cases}$$

where c_i ($i=1, \ldots, 4$) and d_j ($j=1,2$) are constants to be determined for each aggregation rule.

An example diagram presenting $s_{Conf}()$ and $s^{-1}{}_{Conf}()$ is given in Fig. 12.

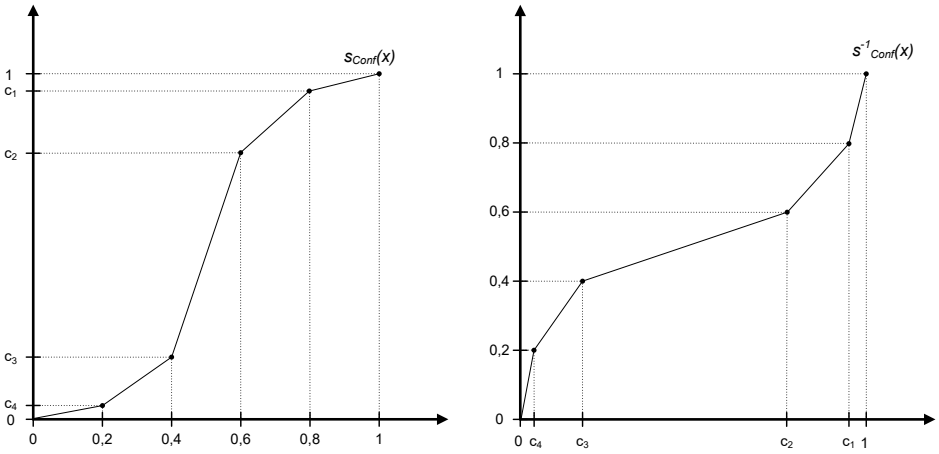

Fig. 12 Example scaling function (Confidence scale for C-rule)

Scaling function $s_{Conf}()$ maps the values which correspond to the levels of the linguistic scale, i.e. *1*, *0.8*, *0.6*, *0.4*, *0.2* and *0* to *1*, c_1, c_2, c_3, c_4 and *0* respectively. The intermediate values are mapped using linear functions, which was adopted as the simplest solution fulfilling the requirement that *s* should be monotonic (the only constraint identified for *s*).

All scaling functions were calculated using the data obtained in the experiment. 31 students studying for the Master's degree in information technologies (the last year) were selected for the experiment. The participants had good background in logic and mathematics and they also attended a two-hour lecture about trust cases.

The participants were divided into three groups. Each group was supposed to apply one of the aggregation rules: A-rule, C-rule or NSC-rule (SC-argument type was dropped because of its similarity to NSC-argument type).

Each participant was provided with five simple trust cases composed of a claim, an argument strategy, a warrant and premises (in the case of C-rule and NSC-rule) or a claim with a few argument strategies (in the case of A-rule). One of the trust cases is presented in Fig. 13.

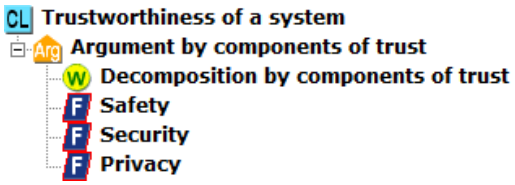

Fig. 13 Example trust case

The experiment participants were asked to assess the warrant and, in the case of C-rule to assign weights to the premises. Then, assuming the pre-defined assessments of each premise (in the case of C-rule and NSC-rule) or the assessments assigned to each of the argument strategies (in the case of A-rule) the participants were asked to

give their assessment of the conclusion using the Assessment Triangle. 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 the total of 50 assessments of the conclusions issued by each participant. To check for consistency, 10 randomly selected assessments were then repeated for each participant.

Some participants were excluded from the experiment for formal reasons (e.g. one of them was cheating while filling the consistency questionnaire in, others did not provide all the answers) 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, which is clearly logically wrong and was significantly different from the answers provided to the same question by the rest of the group). 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 applying the following criteria:

Consistency of assessments - measured by calculating the root-mean-square value of the difference between the first and the repeated assessment (by the same participant) of the same conclusion with the same assessments assigned to the premises.

Accuracy of assessments - measured by calculating the root-mean-square value of the difference between a participant's assessment and the result of application of the aggregation rule.

The above calculations were performed for both, Confidence and Decision scales. The results are presented in Table 1. The numbers are normalized, which means that 1 represents the distance between two adjacent values 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 the participants' answers. This is the maximum of what could have been achieved regarding the data set used to calibrate the aggregation rules.

Table 1 Results of the experiment

Aggregation rule	**Consistency of assessments**		**Accuracy of assessments**	
	Confidence scale	Decision scale	Confidence scale	Decision scale
A-rule	1.03	0.64	1.06	0.80
C-rule	0.94	0.62	1.10	0.78
NSC-rule	0.84	0.87	0.90	0.66

Further calibration requires more data which we plan to collect in the subsequent experiments.

11 Conclusion

This article introduced the Visual Assessment of Arguments (VAA) method which proposed an innovative approach to argument structures appraisal. The method provides for gathering expert opinions about the inferences used in the argumentation and the value of the supporting evidence. It can be applied to assess (with the help of experts) the compelling power of arguments used in different contexts. In particular, it can be used with respect to arguments contained in different types of cases, like safety cases, security cases, assurance cases or trust cases or within the context of the collaborative development process for trust cases [6, 15].

The method has already been fully implemented in the TCT tool which supports full-scale application of the Trust-IT framework [12, 13, 14].

VAA has been subjected to experimental validation and further experiments are under preparation. Furthermore, the method has been applied for appraisal of arguments for patient safety and privacy, and for fulfilment of security requirements in two EU funded 6th FR projects. It is also going to be used in a new project utilising argument structures to demonstrate conformity with standards and regulations (the project is planned to commence from 2010).

Acknowledgement

Contributions by Michal Nawrot and Michal Witkowicz in the development of the appraisal mechanism for the TCT tool are to be acknowledged.

References

[1] ANGEL project website, Advanced Networked embedded platform as a Gateway to Enhance quality of Life, ftp://ftp.cordis.europa.eu/pub/ist/docs/dir_c/ems/angel-v1.pdf.

[2] Bishop P, Bloomfield R. A methodology for safety case development. Industrial Perspectives of Safety-Critical Systems. Proceedings of the Sixth Safety-Critical Systems Symposium, Birmingham; 1998.

[3] Bloomfield R, Guerra S, Masera M, Miller A, Sami Saydjari O. Assurance cases for security. A report from a Workshop on Assurance Cases for Security. Washington, USA; 2005.

[4] Bloomfield RE, Littlewood B, Wright D. Confidence: its role in dependability cases for risk assessment. Proceedings of 37th annual IEEE/IFIP International Conference of Dependable Systems and Networks; 2007, p. 338-346.

[5] Cyra L, A method of trust case templates to support standards conformity achievement and assessment. PhD thesis; 2008; Gdansk, Poland.

[6] Cyra L, Gorski J. Expert assessment of arguments: a method and its experimental evaluation. Proceedings of 27th International Conference on Computer Safety, Reliability and Security SAFECOMP; 2008; Newcastle, UK. Berlin/Heidelberg: Springer, Lecture Notes in Computer Science 5219; 2008, p. 291-304.

[7] Cyra L, Gorski J. Supporting compliance with safety standards by trust case templates. Proceedings of ESREL; 2007; Norway. 2, p. 1367-1374.

[8] Cyra L, Gorski J. Standard compliance framework for effective requirements communication. Polish Journal of Environmental Studies; 2007; 16:5B, p. 312-316.

[9] Cyra L, Gorski J. Extending GQM by argument structures. Proceedings of 2nd IFIP Central and East European Conference on Software Engineering Techniques CEE-SET; 2007, p. 1-16.
[10] Cyra L, Gorski J. Using argument structures to create a measurement plan. Polish Journal of Environmental Studies; 2007; 16:5B, p. 230-4.
[11] Cyra L, Gorski J. Supporting expert assessment of argument structures in trust cases. Proceedings of Ninth International Probabilistic Safety Assessment and Management Conference PSAM 9; 2008; Hong Kong, China, p. 1-9.
[12] Gorski J, Trust-IT - a framework for trust cases. Proceedings of Workshop on Assurance Cases for Security - The Metrics Challenge, The 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks DSN; 2007; Edinburgh, UK. p. 204-9.
[13] Gorski J, Jarzebowicz A, Leszczyna R, Miler J, Olszewski M. Trust Case: Justifying Trust in IT Solution. Elsevier, Reliability Engineering and System Safety; 2005, 89, p. 33-47.
[14] Gorski J. Trust Case – a Case for Trustworthiness of IT Infrastructures. Cyberspace Security and Defence: research issues, NATO ARW series, Springer-Verlag; 2005, p. 125-142.
[15] Gorski J. Collaborative approach to trustworthiness of IT infrastructures. Proceedings of IEEE International Conference on Technologies for Homeland Security and Safety TEHOSS; 2005, p. 137-142.
[16] Gorski J, Zagorski M. Reasoning about trust in IT infrastructures. Proceedings of ESREL; 2005, p. 689-695.
[17] Greenwell W, Knight J, Holloway C, Pease J. A taxonomy of fallacies in system safety arguments. Proceedings of the 24th International System Safety Conference; 2006.
[18] Greenwell W, Strunk E, Knight J. Failure analysis and the safety-case lifecycle. Human Error, Safety and Systems Development; 2004, p. 163-176.
[19] Information Assurance Group. TCT user manual, Gdansk university of technology. http://kio.eti.pg.gda.pl/trust_case/download/TCTEditor_Users_Manual.pdf; 2007.
[20] Josang A, Grandison T. Conditional inference in subjective logic. Proceedings of the 6th International Conference on Information Fusion Cairns; 2003, p. 471-8.
[21] Josang A, Pope S, Daniel M. Conditional deduction under uncertainty. Proceedings of the 8th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty ECSQARU; 2005; Barcelona, Spain. p. 824-835.
[22] Kelly T. Arguing safety – a systematic approach to managing safety cases. PhD thesis; 1998; University of York, UK.
[23] Kelly T, McDermid J. A systematic approach to safety case maintenance. Proceedings of SAFECOMP; 1999, p. 271-284.
[24] Littlewood B, Wright D. The use of multilegged arguments to increase confidence in safety claims for software-based systems: a study based on a BBN analysis of an idealized example. IEEE Transactions on Software Engineering; 2007, p. 347-365.
[25] Mayo P. Structured safety case evaluation: a systematic approach to safety case review. Proceedings of the 1st IET International Conference on System Safety; 2006, p. 164-173.
[26] PIPS project website, Personalised Information Platform for Health and Life Services, http://193.178.235.132:8088/pips.
[27] Sentez K, Ferson S. Combination of evidence in Dempster-Shafer theory. SANDIA National Laboratories; 2002.
[28] Shafer G. Mathematical theory of evidence. Princetown University Press; 1976.
[29] Strigini L. Formalism and judgement in assurance cases. Workshop on Assurance Cases: Best Practices, Possible Obstacles, and Future Opportunities, DSN; 2004; Florence, Italy. 2004.

[30] Toulmin S. The uses of argument. Cambridge University Press, Cambridge; 1969.
[31] Weaver R, Mayor P, Kelly T. Gaining confidence in goal-based safety cases. Proceedings of 14th Safety Critical Systems Symposium; 2006.
[32] Wu W, Kelly T. Combining Bayesian belief networks and the goal structuring notation to support architectural reasoning about safety. Proceedings of SAFECOMP; 2007; p. 172-186.
[33] Zechner A, Huhn M. Analysing dependability case arguments using quality models. Proceedings of 28th International Conference on Computer Safety, Reliability and Security SAFECOMP; 2009.