# A Goal-Based Approach for Qualification of New Technologies: Foundations, Tool Support, and Industrial Validation

Mehrdad Sabetzadeh[a,*], Davide Falessi[b], Lionel Briand[a],
Stefano Di Alesio[a,c]

[a] *University of Luxembourg, Luxembourg*
[b] *Fraunhofer USA, Center for Experimental Software Engineering, USA*
[c] *Simula Research Laboratory, Norway*

## Abstract

New technologies typically involve innovative aspects that are not addressed by the existing normative standards and hence are not assessable through common certification procedures. To ensure that new technologies can be implemented in a safe and reliable manner, a specific kind of assessment is performed, which in many industries, e.g., the energy sector, is known as Technology Qualification (TQ). TQ aims at demonstrating with an acceptable level of confidence that a new technology will function within specified limits. Expert opinion plays an important role in TQ, both to identify the safety and reliability evidence that needs to be developed, and to interpret the evidence provided. Since there are often multiple experts involved in TQ, it is crucial to apply a structured process for eliciting expert opinions, and to use this information systematically when analyzing the satisfaction of a technology's safety and reliability objectives.

In this article, we present a goal-based approach for TQ. Our approach enables analysts to quantitatively reason about the satisfaction of a technology's overall goals and further to identify the aspects that must be improved to increase goal satisfaction. The approach is founded on three main components: goal models, expert elicitation, and probabilistic simulation. We

[*] Corresponding author

*Email addresses:* mehrdad.sabetzadeh@uni.lu (Mehrdad Sabetzadeh), dfalessi@fc-md.umd.edu (Davide Falessi), lionel.briand@uni.lu (Lionel Briand), stefanod@simula.no (Stefano Di Alesio)

describe a tool, named Modus, that we have developed in support of our approach. We provide an extensive empirical validation of our approach through two industrial case studies and a survey.

*Keywords:* Technology Qualification, Assurance Cases, Goal Modeling, Expert Elicitation, Monte Carlo Simulation, Offshore Systems.

## 1. Introduction

Most systems in critical application areas such as healthcare, avionics, and energy are subject to some form of assessment to ensure that the risks associated with the use of the systems are properly mitigated. The most widely-known type of assessment is certification, conducted by an independent professional or regulatory body, to verify that a system is in compliance with one or more applicable standards. In fast-growing markets, such as the energy sector, assessors are frequently faced with innovative technologies that are not fully addressed by the existing standards and hence are not assessable through common certification procedures. To verify that a new technology will work as intended, a specific type of assessment is performed, which in many industries, e.g., the energy sector, is known as Technology Qualification (TQ). Briefly, TQ is aimed at demonstrating with an acceptable level of confidence that a *new* technology will function as intended within specified limits.

To better illustrate the situations where TQ is applied, let us consider two example from the energy and offshore domain:

- ***Marine propulsion.*** For commercial ships, existing propulsion standards are targeted at mechanical propulsion (involving an engine and a propeller). Recently, new technologies have emerged which use kites to harness wind for propulsion, thus reducing the carbon-footprint. While both mechanical and wind propulsion are means to the same end, the principal concepts in mechanical propulsion standards (e.g., engine, propeller, shafts, flywheel, etc.) no longer apply.

- ***Fiber Ropes.*** Steel cables have been used for a long time as the primary apparatus for mooring and installation of floating and underwater structures. Recently, there has been a growing interest in *fiber rope* technologies, both as an alternative to steel cables, and further to enable operations that were previously not possible (e.g., installation

> in deep water). Existing standards for mooring and installation tend to focus on steel cables. Since steel and fiber have very different physical and mechanical properties, these standards are neither fully applicable to, nor cover the entire set of concerns relevant to fiber ropes.

In cases like the above, TQ is instrumental (and sometimes mandatory) to ensure that the new technologies can be deployed in a safe, reliable, and environment-friendly manner.

In this article, building on the notion of goal-based assurance cases [17], we present a quantitative assessment approach for TQ. Our approach, which is supported by a software tool, includes three main components: *goal models*, *expert elicitation*, and *probabilistic simulation*. We use the KAOS goal modeling notation [29] to structure and decompose a technology's (safety and reliability) goals. We apply expert elicitation techniques [1, 19] for soliciting expert probabilities based on the collected evidence and for mitigating potential biases. Arguments about dependability generally have a strong reliance on expert judgment [19]. This is also true in TQ. Dependence on expert judgment is particularly strong in early TQ stages where little evidence exists about a new technology. One of the aims of TQ is to identify critical areas where there is significant uncertainty in expert judgments and to define objective fitness criteria to reduce the uncertainty and dependence on subjective opinions. This all makes it important to follow a rigorous expert elicitation process in TQ. Lastly, we use Monte Carlo simulation [23] to measure goal satisfaction and to identify the weak links that must be improved for reducing the uncertainty in the satisfaction of high-level goals.

Our contributions are: (1) Tailoring expert probability elicitation into (KAOS) goal models; (2) enhancing Requirements Engineering goal propagation methods [8, 10] with Monte Carlo-based analysis; and (3) applying the KAOS notation and our enhanced propagation solution in the context of TQ. The foundations of our approach are general and can be used for various types of assessment, but the methodological steps of our work are motivated by the workflow of activities in TQ. To encourage industrial adoption of our contributions, we align our work with the guidelines in DNV's Recommended Practice for Technology Qualification [22] and Offshore Service Specification [28].

Our approach is supported by a tool, called Modus, which provides features for goal modeling, expert probability elicitation, and probabilistic analysis. We have completed two industrial case studies using the tool. Both case studies are related to assessing the behavior of fiber ropes in safety-critical offshore systems. Overall, the case studies indicate that our approach can support quantitative measurement of goal satisfaction in TQ with a reasonable level of effort.

As a follow-on to the case studies, we have conducted a survey in order to systematically examine the perceptions of the involved experts about our approach. The survey results indicate that our approach offers benefits by improving various quality attributes in TQ, such as productivity and traceability.

The remainder of the article is structured as follows: In Section 2, we give a summary of the TQ process and motivate our work in that context. We describe our approach and its components in Section 3. We discuss tool support in Section 4. We present an empirical evaluation of approach and the insights we have gained through the process in Section 5. In Section 6, we highlight practical considerations and limitations for our approach. We compare our approach with related work in Section 7 and conclude the article in Section 8 with a summary and suggestions for future work.

Parts of this article have been previously published in a research paper at the 13th IEEE International High Assurance Systems Engineering Symposium [24]. This article enhances our earlier work with a more comprehensive description of tool support (Section 4) and substantial new experimental results demonstrating the feasibility and usefulness of our approach (Section 5).

## 2. Background and Motivation

In this section, we provide a brief summary of the activities in TQ (based on DNV RP-A203 [22] and OSS-401 [28]), along with the observations that motivated the development of a goal-based assessment approach for use in this context.

1. ***Specification of Qualification Basis.*** TQ begins with the development of a qualification basis. The basis covers: (1) the technology's main objectives and expectations expressed as functional requirements and environmental parameters, and (2) technical specifications for deployment, operation, and decommissioning of the technology.
2. ***Elaboration of Novel Aspects***. The technology's novel aspects (functions, components, processes) are identified. These aspects are

then decomposed to a level of detail at which potential failure mechanisms can be determined, analyzed, and prioritized. This decomposition is performed by qualified experts representing the relevant technical disciplines and fields of experience.

3. ***Planning and Collection of Evidence.*** An evidence collection plan is developed and the plan is executed. Evidence collection activities are targeted at providing quantitative measures, predominantly in probabilistic terms, for the uncertainties and likelihoods of failure. The evidence can, among others, include laboratory tests, theoretical analyses and simulations, procedural changes to avoid potential problems, and tests to reduce uncertainty in analytical models, e.g., erosion models.
4. ***Verification.*** This involves analyzing the qualification basis, the risk studies, and the collected evidence to confirm that the requirements in the qualification basis are met, and that the identified risks are properly mitigated.

We note that while we present the TQ steps in a linear manner, in practice, TQ is an iterative process. This means that before deployment, the technology concept may undergo several rounds of improvement based on the observations and the results at different steps of the TQ process.

The main motivations for the approach we propose in this article come from observing the current practice in TQ. Specifically, we observed the following issues:

- **A: Traceability and Rationale.** Verification can be challenging as the assessment body must establish that there is a demonstrable link among (1) the qualification basis, (2) the identified risks, and (3) the collected evidence. A compliance matrix can be used to establish these links, but this approach is limited in that it does not record the reasoning as to why different elements are linked.

- **B: Handling of Expert Judgment.** The process taken to elicit expert judgments, the information elicited from the experts, and the way the information is compiled is not always made explicit. This can have a negative impact on the transparency of the TQ process, thus making it hard for both the assessment body and potential end-users to build the required level of trust in the new technology.

- **C: TQ Costs.** Evidence collection (mainly testing) accounts for the majority of TQ costs. Time and budget overruns may occur if effort is not focused on building or improving the right evidence information. Two issues are frequently raised: (1) Vendors undertake costly tests which turn out to be tangential to the TQ process. (2) Vendors undertake appropriate tests but verification indicates a lack of confidence about whether the TQ requirements are met. In such cases, it is difficult to determine which aspects of the evidence need to be improved because the main factors contributing to the uncertainty about the satisfaction of TQ requirements cannot be easily identified.

In our approach, described in Section 3, we use goal models to address **A** by maintaining a logical trace of how TQ requirements are decomposed and linked to the relevant risks and evidence. The decomposition also ensures that evidence collection can be better planned, thus reducing the likelihood of collecting non-useful evidence – see first point of **C** above. To address **B**, an explicit mechanism is incorporated into goal models to elicit, record, and propagate expert probabilities. For the second point in **C**, we use a probabilistic technique known as sensitivity analysis [23] to identify the main sources of uncertainty in the satisfaction of a given goal, thus helping focus TQ resources on providing the evidence that reduces the most important uncertainties first.

## 3. Approach

Figure 1 shows an overview of our approach: we begin with constructing a goal model where we decompose the technology's overall safety and reliability goals, as identified in the technology qualification basis (see Section 2), into more concrete subgoals and obstacles[1].

The next step is devising and executing an evidence collection plan. The evidence enables quantifying the following: (1) probability of low-level goals being satisfied, (2) probability of low-level obstacles occurring, and (3) probability of risks arising from incomplete goal decomposition. In this work, we do not aim to provide a specific solution for evidence planning and collection, as this step largely depends on the technology being assessed [22]. Following

[1] Obstacles are events that obstruct goal satisfaction (see Section 3.1).

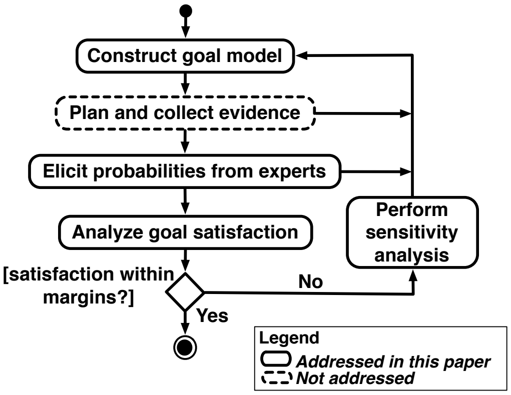

Construct goal model
Plan and collect evidence
Elicit probabilities from experts
Analyze goal satisfaction
Perform sensitivity analysis
[satisfaction within margins?]
No
Yes
Legend
Addressed in this paper
Not addressed

Figure 1: Approach overview.

evidence collection, experts use the evidence to form opinions about the three probability types above.

We then automatically propagate the elicited probabilities to compute probability distributions for the satisfaction of the technology's overall goals. If there is too much uncertainty about the satisfaction of the overall goals, sensitivity analysis will be performed to identify the input quantities with the most significant impact on the uncertainty in goal satisfaction. Based on the results of this analysis, we can go back to the previous steps to make improvements, such as including additional provisions in the technology (leading to goal-model updates), using more dependable components in the technology, collecting further evidence, and using additional or more suitable individuals for expert elicitation. The iterative nature of the TQ process is further justification for an explicit argument model, that can be modified, and then used for re-running the analysis.

### *3.1. Goal Modeling*

We use goal models for decomposing a technology's safety and reliability goals into more specific criteria for which concrete evidence can be collected. Several languages exist for goal modeling; notable examples are: $i^*$ [31], GSN [17], and KAOS [29]. While the main ideas of our approach can be used in conjunction with any of these languages, we choose to ground our work on (a subset of) the KAOS language. This choice is motivated by two main reasons: (1) KAOS formal decomposition semantics is a suitable fit for the type of quantitative reasoning needed in TQ; and (2) KAOS comes with an

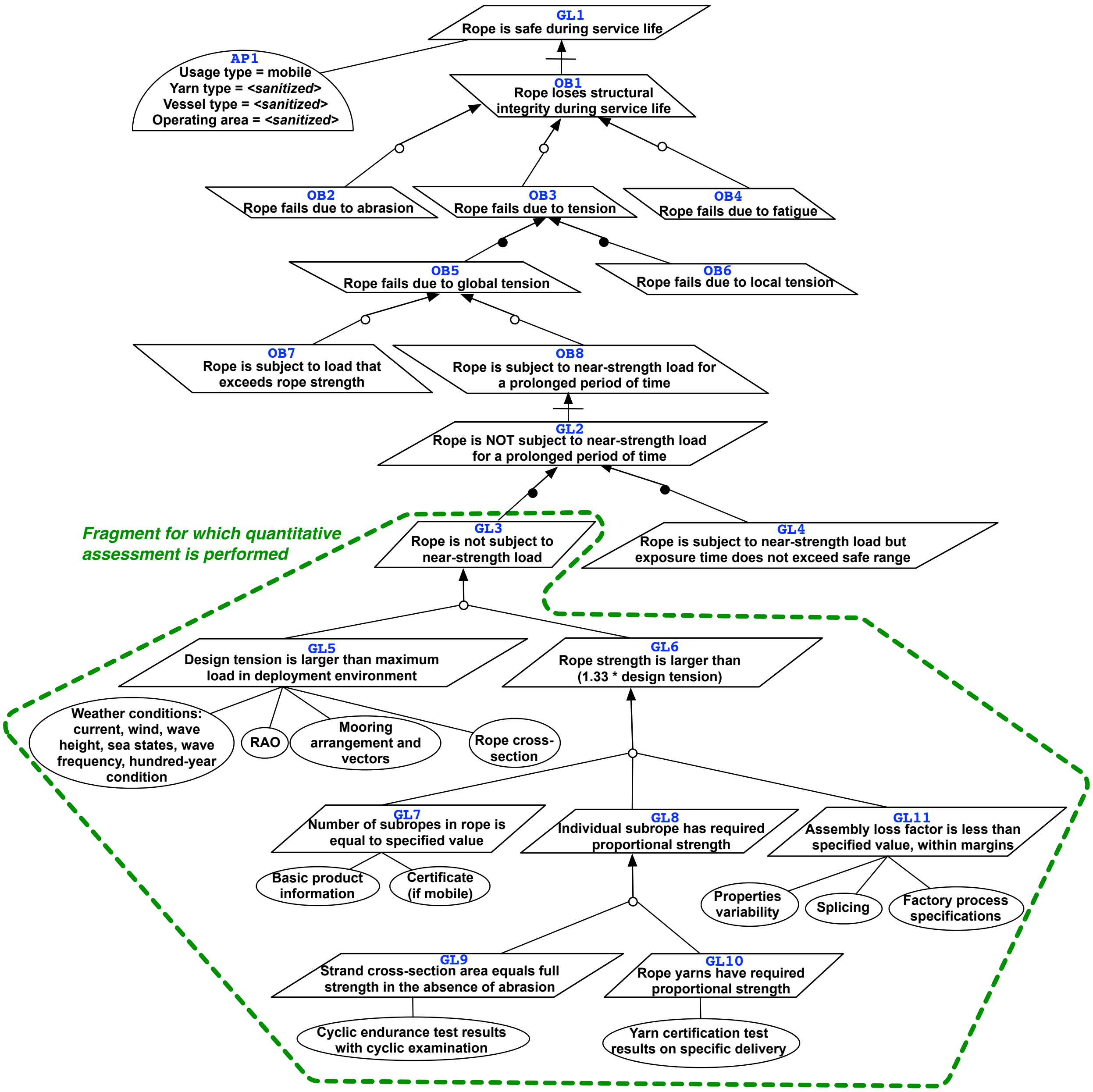

GL1
Rope is safe during service life
AP1
Usage type = mobile
Yarn type = <sanitized>
Vessel type = <sanitized>
Operating area = <sanitized>
OB1
Rope loses structural integrity during service life
OB2
Rope fails due to abrasion
OB3
Rope fails due to tension
OB4
Rope fails due to fatigue
OB5
Rope fails due to global tension
OB6
Rope fails due to local tension
OB7
Rope is subject to load that exceeds rope strength
OB8
Rope is subject to near-strength load for a prolonged period of time
GL2
Rope is NOT subject to near-strength load for a prolonged period of time
Fragment for which quantitative assessment is performed
GL3
Rope is not subject to near-strength load
GL4
Rope is subject to near-strength load but exposure time does not exceed safe range
GL5
Design tension is larger than maximum load in deployment environment
GL6
Rope strength is larger than (1.33 * design tension)
Weather conditions: current, wind, wave height, sea states, wave frequency, hundred-year condition
RAO
Mooring arrangement and vectors
Rope cross-section
GL7
Number of subropes in rope is equal to specified value
GL8
Individual subrope has required proportional strength
GL11
Assembly loss factor is less than specified value, within margins
Basic product information
Certificate (if mobile)
Properties variability
Splicing
Factory process specifications
GL9
Strand cross-section area equals full strength in the absence of abrasion
GL10
Rope yarns have required proportional strength
Cyclic endurance test results with cyclic examination
Yarn certification test results on specific delivery

Figure 2: Simplified goal model for fiber rope mooring.

extended and unified set of modeling guidelines in a book [29]. Such a book is an advantage for training and technology transfer to practitioners.

To illustrate goal modeling in KAOS, we use a simplified and sanitized version of the goal model in one of our industrial case studies (see Section 5.1) for arguing about fiber rope safety in mooring systems. Figure 2 shows this goal model. We concentrate on a particular aspect of the fiber rope behavior that is markedly different from steel chains. In design, the integrity over prolonged time is handled by design curves. The design curve for steel chains describes the safe service life in a relationship between loading range and number of stress cycles. In contrast, the design curve for fiber ropes expresses the relationship between static tension and time to rupture.

In less technical terms, if a steel chain is not subject to cyclic loading (resulting in fatigue), the chain will maintain its integrity for tens of years under a static tension almost equal to its maximum strength. Fiber ropes in contrast exhibit a time-dependent behavior under tension [18]: Under a static tension of up to 65% of its maximum strength, a fiber rope has a safe life (w.r.t. tension failures) of tens of years. However, beyond 65%, the rope's safe life starts to deteriorate as tension increases. For example, if subject to a tension of 75% of its strength, a fiber rope will on average rupture after tens of days. Increasing the tension to 85% of rope strength will decrease the safe life to a few hours. Due to the logarithmic relationship between tension and stress-rupture time for fiber ropes, design considerations have to be made when tension exceeds 75% of the fiber rope strength [21]. This is to ensure that the rope can withstand the longest and most severe storms at sea.

Specifically, what TQ needs to ascertain here is the following: for a specific type of fiber rope used in a specific environment, safety is not compromised as the result of the rope's time-dependent integrity even during a major storm potentially lasting for several days. The model fragment that is the subject of quantitative assessment (Section 3.3) is distinguished in Figure 2 with a dashed boundary.

Each goal is "a prescriptive statement of intent that should be satisfied" [29]. Goals in KAOS are depicted in the parallelogram shape (e.g., `GL1`). The assumptions under which a given goal is to be satisfied are made explicit and captured via an assumption node, presented as a semicircle (e.g., `AP1`). Goal decomposition is performed using AND and OR operators to show either the case where several subgoals together contribute to the satisfaction of the parent goal, or where alternatives exist for satisfaction. The decomposition can be either full or partial. Full decomposition means that a parent goal has been completely refined and that no more subgoals will be added to the decomposition; whereas partial decomposition means that more subgoals may be added in the future. Partial decomposition is shown using an empty circle and full decomposition is shown using a filled circle. In Figure 2, we use both full and partial decomposition. For example, `GL2` is OR-decomposed using full decomposition and `GL3` is AND-decomposed using partial decomposition.

The obstacles that prevent (obstruct) the satisfaction of goals are depicted as mirrored parallelograms. For example, in Figure 2, the obvious obstacle to the fulfillment of `GL1` is that the rope breaks `OB1`. This is called the root obstacle. The root obstacle is then decomposed into the factors that can lead to it (in this case, `OB2`–`OB4`). Obstacle decomposition is done in exactly the same manner as goal decomposition. Further, just as obstacles can obstruct goal satisfaction, goals can mitigate obstacles (e.g., `GL2` mitigates `OB8`).

Goal and obstacle decomposition continues until we reach criteria that are fine-grained enough to be supported by concrete evidence. Evidence items are depicted as ovals and are linked to the relevant leaf goals and obstacles (e.g., see ovals connected to `GL5`). The standard KAOS language does not provide notational elements for representing assumptions and evidence items. The notation we use for evidence items is borrowed from GSN [17].

Developing a goal model before planning the evidence makes the evidence collection process more targeted and helps avoid activities with limited usefulness, e.g., an expensive full-scale test that despite its impressiveness does not challenge the technology at the operational boundaries. The evidence item(s) linked to each leaf goal (or obstacle) provide experts with information that supports the estimation of the probability of goal satisfaction (or obstacle occurrence). These probabilities are then propagated up the goal model to assess if the overall goals are adequately satisfied. We discuss expert elicitation and goal propagation in Sections 3.2 and 3.3 respectively.

### *3.2. Expert Probability Elicitation*

We start this section by describing the probabilistic quantities that we need to elicit from the experts in the TQ process (Section 3.2.1). Drawing on the existing literature [1, 19], we then propose a simple elicitation method for use in TQ (Section 3.2.2), along with a suitable elicitation protocol (Section 3.2.3). Since TQ is targeted at assessing new technologies with a limited operational profile, it has to be able to cope with the uncertainty arising from partial and potentially conflicting evidence information. In particular, due to this uncertainty, the experts may be unable to quantify their probability estimates using exact (point-value) probabilities. Our proposed elicitation solution is therefore intended at the situation where a probability distribution has to be specified for each estimate rather than a point-value. Using distributions as input for the assessment also leads to the assessment results being in terms of distributions (and not point-values). We discuss the computation and interpretation of the assessment results in Section 3.3.

### *3.2.1. Description of Elicitation Quantities*

There are three types of probabilities that need to be elicited from the experts in our approach:

- *Probability of satisfaction of a leaf goal*: Experts provide the probability of a (leaf) goal to be satisfied. In particular, given a (leaf) goal $G$ and supporting evidence items $E_1, \ldots, E_\ell$, experts need to answer the following: "Based on $E_1, \ldots, E_\ell$, how likely is $G$ to be satisfied?"

- *Probability of occurrence of a leaf obstacle*: Given a (leaf) obstacle $O$ and supporting evidence items $E_1, \ldots, E_\ell$, experts need to answer the following: "Based on $E_1, \ldots, E_\ell$, how likely is $O$ to occur?"

- *Probability of incompleteness risks*: When a goal or obstacle is decomposed, the decomposition may be partial (see Section 3.1). From a risk assessment perspective, it is reasonable to treat partial *OR* decomposition for *goals* and partial *AND* decomposition for *obstacles* as complete because these kinds of partiality do not impose hidden risks. For example, in the case of partial OR for goals, we are not interested in the probability that a parent goal is satisfied although none of its OR-children have been satisfied. The same applies to partial AND-decomposition of obstacles. However, both partial AND-decomposition for goals and partial OR-decomposition for obstacles pose risks. Therefore, given a parent goal $G$ (resp. obstacle $O$) and subgoals $G_1, \ldots G_n$ (resp. sub-obstacles $O_1, \ldots, O_n$), the experts need to answer the following:

  - *Partial AND for goals*: "How likely is goal $G$ to fail despite all subgoals $G_1, \ldots, G_n$ being satisfied?". We denote the answer by $\alpha$.
  - *Partial OR for obstacles*: "How likely is obstacle $O$ to occur despite none of sub-obstacles $O_1, \ldots, O_n$ having occurred?" We denote the answer by $\beta$.

> We note that, just like for leaf goals and leaf obstacles, one can link partial decompositions to evidence items to support the elicitation of $\alpha$ and $\beta$. In our case studies, we do not have such evidence items; hence, this possibility is not exemplified in the goal model of Figure 2.

For example, to reason about the satisfaction of `GL3` in Figure 2, we need to elicit eight quantities: the probabilities of satisfaction for `GL5`, `GL7`, `GL9`, `GL10`, `GL11`, and the value of $\alpha$ for the decompositions of `GL3`, `GL6`, and `GL8`. We describe the protocol for conducting the elicitation next.

### *3.2.2. Elicitation of Probability Distributions*

As we stated earlier, we are interested in eliciting probability distributions from experts so that we can account for the experts' uncertainty. In our context, the elicitation quantities themselves are probabilities (Section 3.2.1). We are therefore concerned with the elicitation of *probability distributions over probabilities*. Such distributions have for a long time been studied under the heading of *imprecise probabilities* [30]. The simplest generalization of point-value probabilities into imprecise probabilities is to replace each point-value with an interval $[a, b]$ where $a$ and $b$ respectively represent the minimum and maximum probabilities for the quantity being elicited. For example, instead of giving an exact probability, say $10^{-3}$, for the occurrence of an obstacle, one can specify a range, say $[10^{-4}, 10^{-2}]$, to capture the uncertainty. The special case where $a = b$ coincides with a point-value probability.

The main limitation of the interval method is that it gives equal weight to all the values within the interval. That is, values close to the minimum and the maximum are as likely as any other value within the interval. This method is therefore unable to capture the intuition that one often wants to express with an imprecise probability in safety and reliability assessment: that the minimum and maximum are *extremes* with little weight, and that the probability weight should be mainly apportioned to the intermediate values of the interval.

A simple extension to the interval method that captures the above intuition is a *triangular distribution* [1]. To define a triangular distribution, we need three parameters (instead of two for the interval method): the minimum ($a$), the maximum ($b$), and the most likely value ($m$). Because humans tend to underestimate the maximum and overestimate the minimum, the elicited triangular distributions are often adjusted by enlarging the confidence interval [7]. A straightforward approach for adjusting the

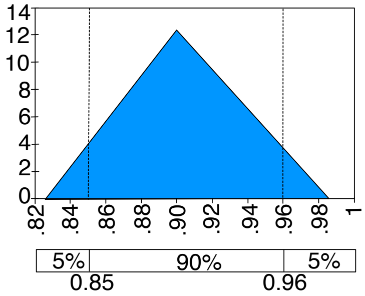


Figure 3: Triangular distribution with extended confidence intervals.

triangular distribution is to extend both end points by equally distributing a certain percentage of the distribution before the minimum and after the maximum [7]. Figure 3 provides an example of triangular distribution with the end points extended by 5% in each direction. The values provided by the expert here are $a = 0.85$, $m = 0.90$, $b = 0.96$. By distributing 5% of the mass around the ends, we get a triangular distribution with $a = 0.82548$, $m = 0.90$, $b = 0.9863$.

A triangular distribution is a good compromise between a crude interval specification of uncertainty and more complex probability distributions such as normal, log-normal, and beta. More complex distributions are both harder to elicit [1] and also more difficult to justify for TQ, as in this context, due to the novelty of the systems being assessed, the exact characteristics of the probability distributions are often unknown; that is, we seldom know for sure if the distribution we aim to elicit is normal, log-normal, beta, etc.

### *3.2.3. Elicitation Protocol*

Our elicitation protocol is based on the guidelines in [19] and consists of the four main steps described below.

- *Step 1 (Recording Experts' Information)* For each expert, we record: name and contact information, field of work, degree, and years of experience.

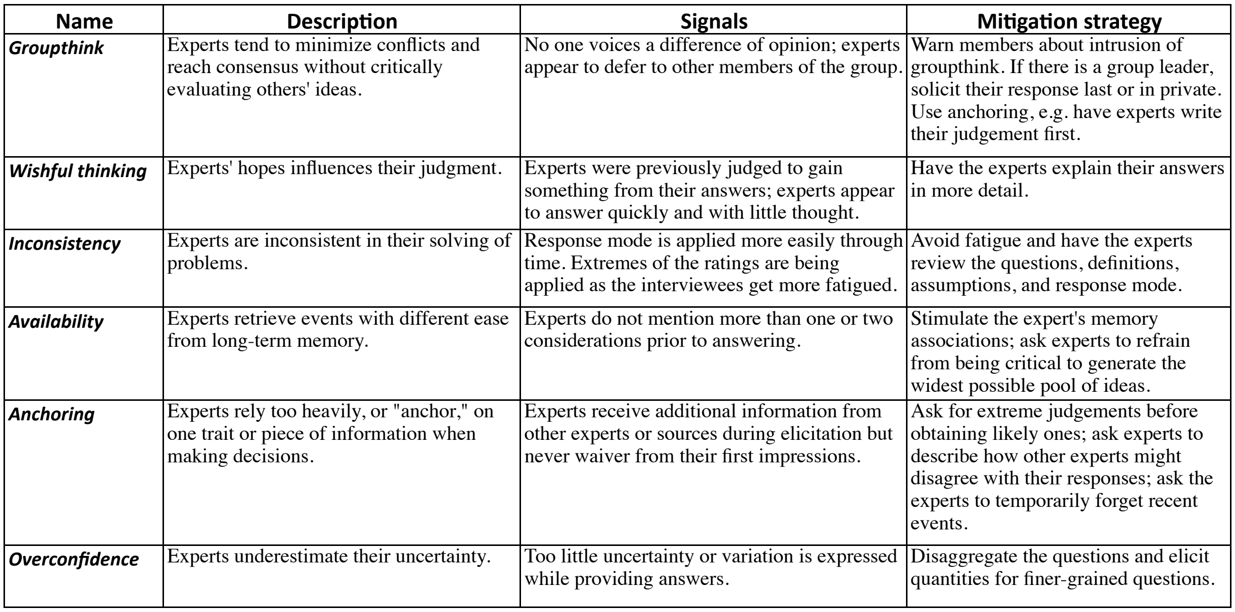
- *Step 2 (Introduction)* We brief experts about the aim of the elicitation, and prepare them by showing examples and highlighting common mistakes. We further make experts aware of the potential intrusion of bias, see Step 4 below.

- *Step 3 (Soliciting answers)* For each quantity, the following process is used:
    - Read the description of the quantity to be elicited.
    - Have experts: (1) explain their understanding of the relevant evidence (or rationale), (2) recall the possible operating conditions and circumstances under which the quantity can be assessed, and (3) relate the setting under which the evidence was collected and the situations arising in practice.
    - Ask: "From your experience and based on the existing evidence, in which operating conditions and circumstances would the probability be very high? What would then be the max. probability?"
    - Ask: "From your experience and based on the existing evidence, in which operating conditions and circumstances would the probability be very low? What would then be the min. probability?"
    - Ask: "Is the most likely value closer to the min. or the max.? What would you deem the most likely value to be?"
    - Define a triangular distribution, based on the elicited max., min., and most likely values, extending the max. and min. values by a pre-specified confidence interval (see Figure 3). As suggested in [7], we used 5% for extending the interval.
- *Step 4 (Handling bias)* During elicitation, the interviewer should monitor the experts' verbalized thoughts and body language, as well as the group dynamics (if a group setting is used for elicitation) for signs of bias. Table 1 summarizes the biases most relevant to TQ, along with mitigation strategies. For a more thorough overview of elicitation biases, consult [19].

### *3.3. Analyzing Goal Satisfaction*

We propagate the values obtained through expert elicitation to compute a satisfaction distribution for each of the overall goals. We describe goal propagation in two steps. First, we provide an algorithm for propagation of point-value probabilities, and then show how we can propagate probability distributions using Monte Carlo simulation.

Table 1: Bias monitoring and mitigation guidelines.



| Name | Description | Signals | Mitigation strategy |
|---|---|---|---|
| ***Groupthink*** | Experts tend to minimize conflicts and reach consensus without critically evaluating others' ideas. | No one voices a difference of opinion; experts appear to defer to other members of the group. | Warn members about intrusion of groupthink. If there is a group leader, solicit their response last or in private. Use anchoring, e.g. have experts write their judgement first. |
| ***Wishful thinking*** | Experts' hopes influences their judgment. | Experts were previously judged to gain something from their answers; experts appear to answer quickly and with little thought. | Have the experts explain their answers in more detail. |
| ***Inconsistency*** | Experts are inconsistent in their solving of problems. | Response mode is applied more easily through time. Extremes of the ratings are being applied as the interviewees get more fatigued. | Avoid fatigue and have the experts review the questions, definitions, assumptions, and response mode. |
| ***Availability*** | Experts retrieve events with different ease from long-term memory. | Experts do not mention more than one or two considerations prior to answering. | Stimulate the expert's memory associations; ask experts to refrain from being critical to generate the widest possible pool of ideas. |
| ***Anchoring*** | Experts rely too heavily, or "anchor," on one trait or piece of information when making decisions. | Experts receive additional information from other experts or sources during elicitation but never waiver from their first impressions. | Ask for extreme judgements before obtaining likely ones; ask experts to describe how other experts might disagree with their responses; ask the experts to temporarily forget recent events. |
| ***Overconfidence*** | Experts underestimate their uncertainty. | Too little uncertainty or variation is expressed while providing answers. | Disaggregate the questions and elicit quantities for finer-grained questions. |

### *3.3.1. Propagation of Point-Value Probabilities*

The basis for propagating point-value probabilities in our approach is the algorithm proposed in [8]. Similar algorithms exist for probability propagation in fault trees [6]. We characterize propagation through the rules shown in Figure 4. In the figure, $P(G_i)$ denotes a (point-value) probability of satisfaction for a goal $G_i$ and $P(O_i)$ denotes a (point-value) probability of occurrence for an obstacle $O_i$. The $\alpha$ and $\beta$ values in rules (b) and (d) are described earlier in Section 3.2. Rules (a)–(d) concern goal–goal propagation. These apply also for obstacle–obstacle propagation; hence, in Figure 4, we do not repeat the rules for obstacles. Rule (e) deals with propagation from a root obstacle to a goal; a dual rule is applied for propagating from a root mitigating goal to an obstacle.

As we also stated in Section 3.2, for goal–goal propagation, we do not need to elicit $\beta$, thus $\beta$ is set to zero and rule (d) reduces to (c) for goals. In contrast, for obstacle-obstacle propagation, $\beta$ is important, whereas $\alpha$ is not, hence reducing rule (b) to (a) for obstacles. Lastly, we note that for rules (c) and (d) to apply, all $G_1, \ldots, G_n$ must be simultaneously realized by the system under assessment as alternative ways to satisfy $G$. In other words, if OR decomposition is used for exploring different alternatives and choosing

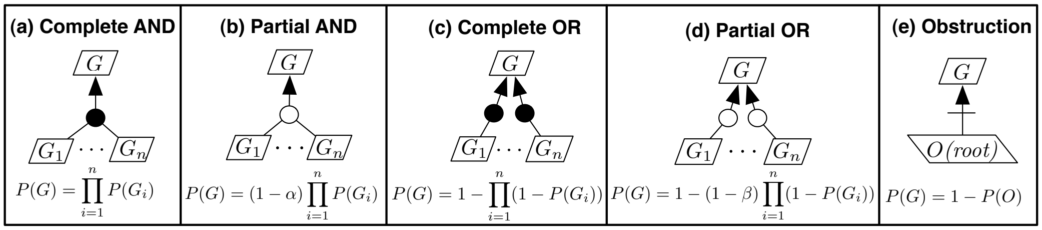


Figure 4: Rules for point-value probability propagation.

one (or a subset) of them for realization, then all the unrealized alternatives must be removed before rules (c) and (d) can be applied.

An assumption underlying goal propagation is that the subgoals that a parent goal are elaborated into are independent from one another. This assumption is common and consistent with best practice in argumentation, which is to decouple arguments that are not explicitly related [8]. As for obstacles, some may represent common-cause failures [6], e.g., loss of electrical power, flooding, ventilation, and human errors. In such cases, an obstacle can obstruct multiple goals. There are several approaches for expressing and quantitatively reasoning about common-cause failures [20]. However, these approaches do not apply at the level of abstraction of goal models, as they require detailed information about the sequence and timing of the failures. In our work, we follow the standard approach in fault-tree analysis [6] and include multiple copies of common-cause obstacles in the goal model. In other words, for a common-cause obstacle $O$, we include a separate copy of $O$ at every location where $O$ is causing an obstruction.

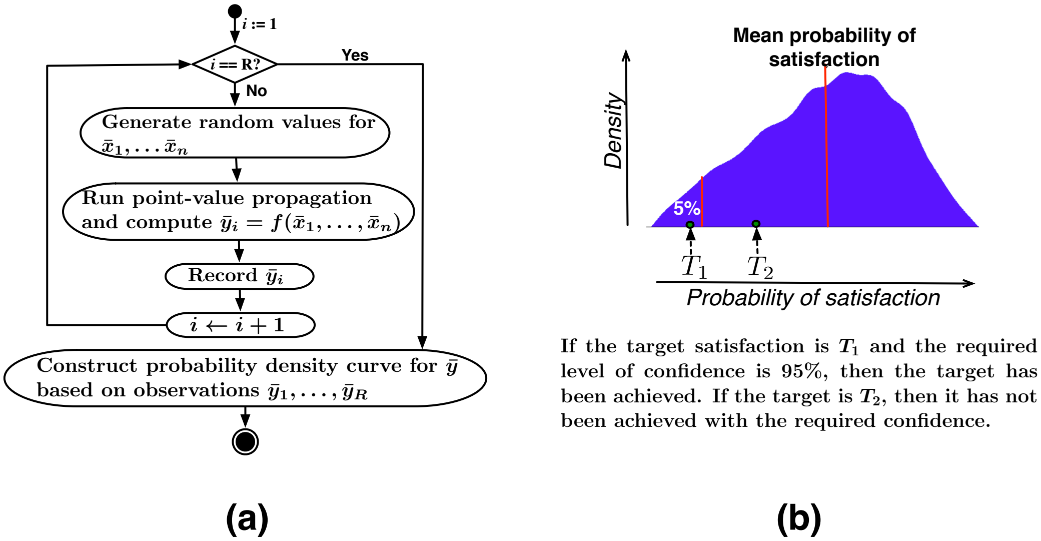
In the random sampling stage of the Monte Carlo simulation (described later in this section), we make sure that in each iteration, only a single random value is drawn for each common-cause obstacle $O$, and not different values for the different instances of $O$ in the goal model. Given the rules in Figure 4, point-value propagation for the entire goal model is performed using the algorithm shown in Figure 5.

### *3.3.2. Propagation of Probability Distributions*

To compute a probability density curve for the satisfaction of an overall goal or the occurrence of an overall obstacle, we use Monte Carlo simulation [23]. The simulation algorithm is shown in Figure 6(a). To run the algorithm, we need to specify the number of iterations ($R$). Each iteration begins with the generation of random input variables ($\bar{x}_1, \ldots, \bar{x_n}$) according

```
▷ Mark all leaf goals and obstacles as "done";
▷ while (tree root is not "done"):
    ▷ for every goal or obstacle $X$ all of whose children are "done":
        ▷ Apply appropriate propagation rule from Figure 4;
        ▷ Mark $X$ as "done";
    ▷ end for;
▷ end while;
▷ output value for tree root;
```

Figure 5: Point-value propagation algorithm.


$i := 1$
$i == R?$
Yes
No
Generate random values for $\bar{x}_1, \ldots \bar{x}_n$
Run point-value propagation and compute $\bar{y}_i = f(\bar{x}_1, \ldots, \bar{x}_n)$
Record $\bar{y}_i$
$i \leftarrow i + 1$
Construct probability density curve for $\bar{y}$ based on observations $\bar{y}_1, \ldots, \bar{y}_R$
Mean probability of satisfaction
Density
5%
$T_1$
$T_2$
Probability of satisfaction
(a)
(b)

**If the target satisfaction is $T_1$ and the required level of confidence is 95%, then the target has been achieved. If the target is $T_2$, then it has not been achieved with the required confidence.**

Figure 6: (a) Monte Carlo algorithm (b) Probability density curve for `GL3` in Figure 2

to the probability distribution for each variable. In our case, the probability distributions for the variables are the triangular distributions elicited from the experts. In the next step, we run the point-value propagation algorithm in Figure 5 and record the resulting value $\bar{y}_i$. After running the algorithm for $R$ rounds, we construct a probability density curve for $\bar{y}$ by computing the frequency of the observed values falling into the different value ranges between the min. and max. observed values for $\bar{y}$.

Figure 6(b) shows the probability density curve for goal GL3, with $R = 10,000$. The curve is based on the actual distributions elicited from the experts for GL5, GL7, GL9, GL10, GL11; and $\alpha$ for the decompositions of GL3, GL6, and GL8. We note that, while this could be different in other situations, the experts in our study provided point-value probabilities for $\alpha$ in all three cases. For privacy, we do not provide the actual quantities elicited from the experts. Further, in Figure 6(b), we report the curve shape without the actual numbers. In addition to showing the mean probability of satisfaction, the curve provides the level of confidence for the satisfaction. To interpret this curve, the analysts can apply the following procedure:

1. Specify the target probability of satisfaction, $0 \leq T \leq 1$, to be achieved for a high-level goal;
2. Specify the level of confidence, $0 \leq C \leq 1$, required for the analysis. The value $1 - C$ denotes the margin of error that can be tolerated; i.e., the risk that the analysis may find the goal as satisfied at level $\geq T$; whereas, the actual satisfaction level is $< T$.
3. Measure the surface $S$ under the curve for the interval $[T, 1]$. If $S \geq C$, then the goal is satisfied; otherwise, the goal is not satisfied.

For example, in Figure 6(b), if the targeted probability of satisfaction for GL3 is $T_1$ and the required level of confidence is 95%, the curve tells us that, based on the existing evidence, the technology fulfills the target within the desired confidence interval. In contrast, the technology does not fulfill $T_2$ within a confidence interval of 95%.

To reduce the uncertainty associated with the satisfaction of a goal, it is important to be able to identify the main factors that contribute to the uncertainty. This is achieved through sensitivity analysis, as discussed next.

### *3.4. Sensitivity Analysis*

Sensitivity analysis is concerned with understanding how the uncertainty in the output of a model can be attributed to different sources of uncertainty in the model input [5]. An input impacts the output significantly if the input has a high variance and this variance is propagated through the model to the output. The uncertainty in an input with high sensitivity can bring about a large impact on the output. In contrast, even a large degree of uncertainty in an input with low sensitivity may have negligible impact on the output. Since there can be many inputs with different degrees of uncertainty that simultaneously affect the output, it is important to be able to determine which inputs are the most sensitive and take remedial measures to reduce their uncertainty.

As suggested by the description above, there are three basic abstractions in sensitivity analysis: the inputs, the (propagation) model, and the output. In our context, the inputs are the probability distributions elicited from the experts, the model is a goal model equipped with the propagation rules discussed in Section 3.3, and the output is the probability distribution computed via Monte Carlo simulation for the top-level goal in the model. Here, sensitivity analysis provides a way to identify and rank the expert probabilities that contribute most to the uncertainty in goal satisfaction. The resulting ranking is a useful guide for the experts to determine the estimates whose uncertainty needs to be reduced through further investigation and perhaps by collecting further evidence.

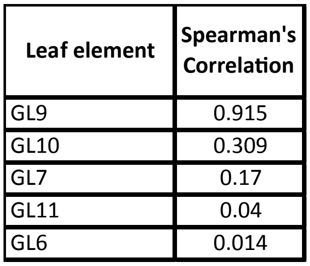
In a probabilistic model, sensitivity analysis is typically conducted using correlation coefficients between inputs and the output [5]. Two of the most commonly used correlation coefficients are: (1) Pearson's correlation, denoted $r$, or (2) Spearman's rank correlation, denoted $\rho$. These measures, which are computed for each input quantity, indicate how sensitive the output is to that particular input distribution.

For Pearson's correlation, quantities should be normally distributed and on interval or ratio scale. Instead, Spearman's rank is non-parametric and only requires ordinal scale. This makes the Spearman's rank more suitable in our context. The higher the correlation between an input and the output, the more significant the influence of the input is on the uncertainty in the output. For example, for `GL3` in Figure 2, it is possible to prioritize the leaf goals based on their impact. In Table 2, we show a list of the leaf goals, sorted by their Spearman's rank computed according to the samples produced by Monte Carlo simulation. Once the leaf goals have been prioritized, one must strive to reduce, as much as possible, the uncertainty in the leafs with the largest correlations.

## 4. Tool Support

We have developed a tool named Modus to support our approach. This tool has been already used with success in the two case studies. Briefly, Modus enables users to (1) construct goal models using the KAOS notation and check the models' structural consistency, (2) link and navigate heterogeneous evidence artifacts, (3) perform the expert elicitation steps of our

Table 2: Sensitivity Results



| Leaf element | Spearman's Correlation |
|---|---|
| GL9 | 0.915 |
| GL10 | 0.309 |
| GL7 | 0.17 |
| GL11 | 0.04 |
| GL6 | 0.014 |

approach and record the elicited probabilities, and (4) export the elicited probabilities and the goal propagation rules as a spreadsheet that can be used for push-button Monte-Carlo simulation and sensitivity analysis.

Modus is implemented as a plug-in for the Enterprise Architect (EA) tool (`http://www.sparxsystems.com/ea`). Among the existing alternatives for modeling environments, we selected EA primarily because of its high usability, widespread use in the industry, availability of detailed instructions for plugin development, and support for storage and linking of heterogeneous evidence information (e.g., requirements and design documents, process descriptions, source code, V&V specifications and results) to goal models.

The tool is written in C# and XML and is approximately 6,000 lines of code. We used Microsoft .NET Framework 2.0 and Visual Studio 2008 as the development platform. A detailed video demonstration of Modus can be found at `http://modelme.simula.no/Modus`.

Figure 7 shows the overall architecture of Modus. All the information related to a given project is stored by EA in a database. Modus can read from and write to this database through the EA's API. In the remainder of this section, we will describe the main components of the Modus plug-in.

### *4.1. Goal Modeler*

Modus makes the notational elements of KAOS available through a toolbox. Figure 8 shows this toolbox (left side) along with a small goal model fragment for a train system (borrowed from [29]). As seen in the figure, all elements are annotated with appropriate labels denoting their types. For example, the top goal ("Stop the train at the STOP signal") has the `goal` annotation and the obstacle that obstructs this goal ("Train not stopped at the STOP signal") has the `obstacle` annotation. The user has the option to hide these annotations if they so wish.

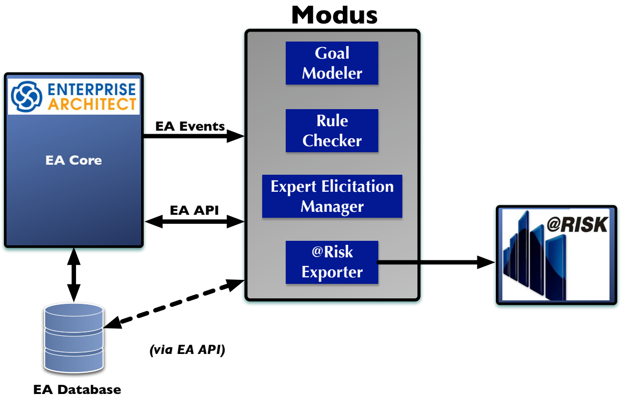
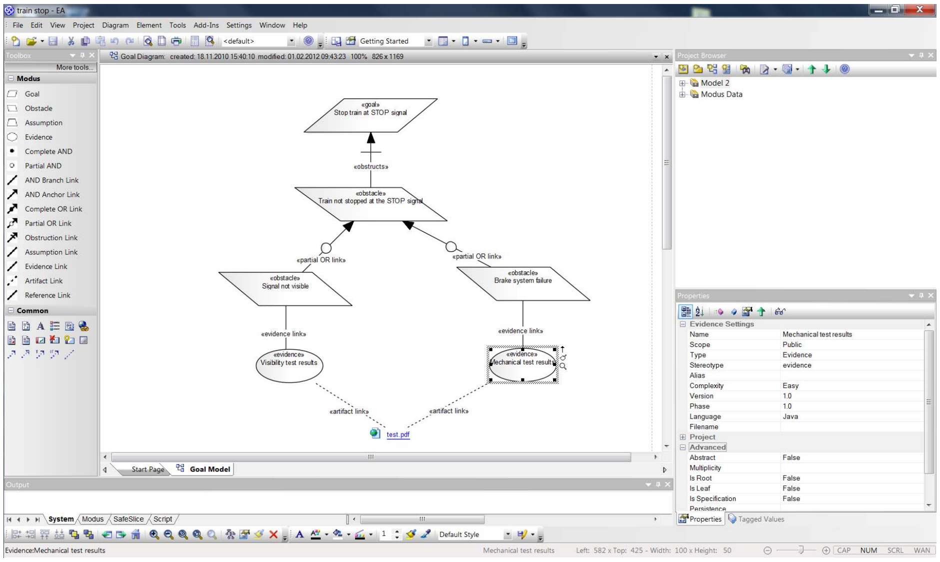


Figure 7: Modus tool architecture



Goal Diagram: created: 18.11.2010 15:40:10 modified: 01.02.2012 09:43:23 100% 826 x 1169

Evidence Settings

| | |
|---|---|
| Name | Mechanical test results |
| Scope | Public |
| Type | Evidence |
| Stereotype | evidence |
| Alias | |
| Complexity | Easy |
| Version | 1.0 |
| Phase | 1.0 |
| Language | Java |
| Filename | |

Project

Advanced

| | |
|---|---|
| Abstract | False |
| Multiplicity | |
| Is Root | False |
| Is Leaf | False |
| Is Specification | False |
| Persistence | |

Evidence:Mechanical test results Mechanical test results Left: 582 x Top: 425 - Width: 100 x Height: 50

Figure 8: KAOS toolbox and example goal model in Modus

As we stated in Section 3.1, our formalism augments KAOS with an explicit way to model evidence information. Specifically, Modus can capture two notions of evidence: (1) *physical evidence*, representing physical artifacts, e.g., test result reports and historical data. Physical evidence is depicted as file links, e.g., `test.pdf` in Figure 8; and, (2) *logical evidence*, providing a logical view on physical evidence. This enables analysts to state what aspects of a (potentially large) physical artifact are relevant to a particular goal or obstacle. Logical evidence is depicted using ovals. Modus supports saving of the physical artifacts (e.g,. Word, Excel or PDF documents) alongside the goal model. This makes it possible to create a *complete* TQ project in one single project file.

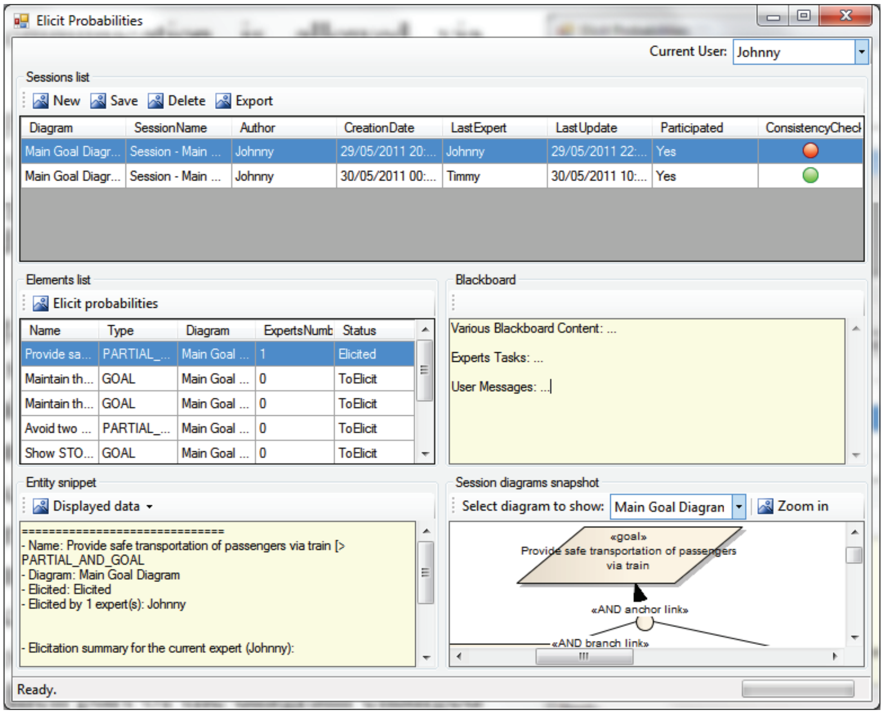
### *4.2. Rule Checker*

To assist users during goal model construction, Modus provides a feature to check the conformance of goal models to the well-formedness rules of the KAOS notation. An example well-formed rule is that an obstacle can obstruct only goals (but not other obstacles), and that a goal can mitigate only obstacles (but not other goals). Modus implements a total of 39 such rules and verifies them using EA's built-in rule engine.

### *4.3. Expert Elicitation Manager*

The Expert Elicitation Manager implements the features related to expert elicitation. Every Modus project includes one or more expert elicitation sessions. Each session keeps track of the experts involved in a given round of elicitation activities and the probabilities provided by the experts. Modus further provides a blackboard feature allowing the experts in a session to interact via exchanging short messages. To obtain probability estimates from the experts, Modus lists all the relevant evidence items for each goal or obstacle, and makes the list available to the experts. To avoid anchoring and maintain confidentiality (which is required in a Delphi setting [19]), experts can choose to make their answers invisible to others.

A screenshot of the expert elicitation interface is shown in Figure 9. On the top, the experts can see a list of elicitation sessions that they are involved in. The center-left of the panel shows the entities whose probabilities are being elicited. If the user double-clicks on an entity, a dialog box will open (not shown). Through this dialog box the user can answer the elicitation questions (discussed in Section 3.2.3), or revise their earlier answers. The center-right of the panel shows the expert communication blackboard. The bottom-left of the panel shows the answers provided by other experts for the entity that has been selected. These answers are visible only if the user chooses to see them, and the other experts have made their answers visible. The bottom-right shows the goal model related to the selected elicitation session.



| Diagram | SessionName | Author | CreationDate | LastExpert | LastUpdate | Participated | ConsistencyCheck |
|---|---|---|---|---|---|---|---|
| Main Goal Diagr... | Session - Main ... | Johnny | 29/05/2011 20:... | Johnny | 29/05/2011 22:... | Yes | 🟥 |
| Main Goal Diagr... | Session - Main ... | Johnny | 30/05/2011 00:... | Timmy | 30/05/2011 10:... | Yes | 🟩 |

| Name | Type | Diagram | ExpertsNumb | Status |
|---|---|---|---|---|
| Provide sa... | PARTIAL_... | Main Goal ... | 1 | Elicited |
| Maintain th... | GOAL | Main Goal ... | 0 | ToElicit |
| Maintain th... | GOAL | Main Goal ... | 0 | ToElicit |
| Avoid two ... | PARTIAL_... | Main Goal ... | 0 | ToElicit |
| Show STO... | GOAL | Main Goal ... | 0 | ToElicit |

Various Blackboard Content: ...

Experts Tasks: ...

User Messages: ...

```
==============================
- Name: Provide safe transportation of passengers via train [>
PARTIAL_AND_GOAL
- Diagram: Main Goal Diagram
- Elicited: Elicited
- Elicited by 1 expert(s): Johnny


- Elicitation summary for the current expert (Johnny):
```

Session diagrams snapshot
Select diagram to show: Main Goal Diagran
Zoom in
«goal»
Provide safe transportation of passengers via train
«AND anchor link»
«AND branch link»
Ready.

Figure 9: Expert elicitation interface

### 4.4. @Risk Exporter

For quantitative assessment, Modus translates goal models (and the related distributions) to a risk management tool called @Risk (`http://www.palisade.com/risk/`). Within @Risk it is easy to run Monte Carlo simulation and sensitivity analysis, and to analyze the resulting distribution for the satisfaction of safety and reliability goals. A sanitized example of the output curve from @Risk was shown earlier in Figure 6.

## 5. Evaluation

In this section, we first report on two case studies that we have performed in the offshore domain. We then describe a survey conducted among the domain experts to understand their perceptions about the approach. The case studies aim at ensuring that the approach is feasible and can be applied with reasonable effort, and the survey – at whether the approach offers benefits if it is adopted into the technology qualification practice.

### *5.1. Industrial Case Studies*

We have conducted two case studies in an industrial setting with the goal of investigating the feasibility of our approach and the level of effort required for its application. Both case studies concentrated on the behavior of fiber ropes but focused on different types of systems: mooring and installation. Fragments of the mooring case were used in earlier sections of the article for illustration.

#### *5.1.1. Context*

Our case studies were conducted in the Technology Qualification Service Line at Det Norske Veritas (DNV). DNV is a notified body[2] specializing in providing risk and conformity assessment services, including, among others, classification, certification, and technology qualification. The Technology Qualification Service Line at DNV engages in a various qualification projects, with a focus on the energy sector, particularly offshore platforms and subsea control systems.

Our first case study considers fiber ropes in offshore mooring and the second – in offshore installation. Briefly, mooring refers to securing a floating structure (e.g., an oil rig) in a fixed location. In this context, ropes are used for attaching the floating structure to poles and anchors. Mooring represents a *static* use of the rope in the sense that, once deployed, the rope is not manipulated by the dynamic operations of the floating structure. Installation, in contrast, is a *dynamic* operation, where the rope is used for lowering and retrieving payloads from the seabed. In this context, the rope interacts with various sheaves and spoolers as it is being lowered into and raised from the water. The two case studies provide complementary perspectives on the safety of fiber ropes, in turn, helping us better examine our proposed assessment approach.

#### *5.1.2. Research Questions*

Our case studies examine the following research questions:

- **RQ1. Is our approach feasible?** More specifically, this question is concerned with: (1) whether it is feasible to elaborate safety and reliability requirements in TQ using KAOS goal models, and (2) whether

[2] A notified body is an independent body appointed by an agency (typically governmental) within a European country as being capable of performing the duties of a notified body as defined by the directives.

it is feasible for the experts to provide quantitative values for the leaf goals and obstacles in a goal tree based on the evidence items linked to them.

- **RQ2. Is the effort involved in the application of our TQ approach acceptable?** The answer to this question will be based on the level of effort spent in the case studies. Effort is an important factor for the successful introduction of a new approach. Unless the experts find the level of effort to be reasonable and commensurate with the scale of a project, it is unlikely that the approach will be adopted.

### *5.1.3. Case Study Selection and Process*

The choice of fiber ropes for our case studies was driven by two main factors: the scale and the availability of experts. The former factor was important because the case studies were conceived as *pilot* investigations. As with most new approaches, it was necessary to build up sufficient evidence about the usefulness of our approach before it could be deployed in any large scale projects. While both of our case studies were performed in real settings and involved real experts, we were limited to repeating the assessment of a technology that had already undergone qualification based on the current TQ practices. The ongoing qualification projects at the time were all large-scale and required resources beyond what was available. On the flip side, the replication of a previous assessment enabled the experts to more easily relate and compare the new and existing approaches in the survey that followed the case studies (Section 5.2). The availability of experts was an important factor as well, because doing any meaningful evaluation of our work required access to specialists in the technology being assessed for the duration of the case studies.

The process taken for each case study closely followed the methodology depicted in Figure 1. In the first step, a goal model was built and validated to express the various safety considerations for fiber ropes. In the second step, the leaf goals and obstacles in the goal tree were linked to the supporting evidence. In our case studies, we did not attempt to construct new evidence items and instead used pre-existing items from the qualifications that had been performed previously. In the second step, we performed expert elicitation on the leaf goals and obstacles as well as the partial decomposition nodes. Expert elicitation was followed by Monte Carlo simulation and sensitivity analysis.

### *5.1.4. Data Collection Methods*

Data collection involved two main activities: (1) goal model construction and (2) expert probability elicitation. In both case studies, these two activities were conducted by five domain experts with background in fiber rope qualification and four researchers. The researchers acted as facilitators. Before the case studies were initiated, a half-day training seminar was held where the researcher group introduced the expert group to goal modeling using KAOS and expert probability elicitation. Below, we describe the processes we used in the two case studies for goal modeling and expert elicitation:

**Constructing a Goal Model.** Goal model construction in the mooring case study started with a read-through, performed by the researchers, of an industrial technical report [18] describing the time-dependent behavior of fiber ropes in mooring systems. The purpose of the read-through was to develop a high-level goal decomposition which could then be refined by the experts. The construction of a preliminary goal model in advance was necessary because the experts had no prior experience with goal modeling. For the subsequent refinement step, we used the goal modeling heuristics proposed in [29]. We frequently used the following heuristics:

- *Refinement through HOW questions*: Subgoals (resp. sub-obstacles) of $G$ ($O$) are found by asking questions such as: "How can goal $G$ be satisfied?" (resp. "How can obstacle $O$ be brought about?").
- *Goal and obstacle negation*: Obstructions (resp. mitigations) were found by negating goals (resp. obstacles) and asking HOW questions.

In the installation case study, which began after the completion of the mooring case study, no high-level goal model was built in advance by the researchers. Instead, the experts were offered assistance to construct the goal models on their own. The assistance was in relation to the following: (1) proper use of the KAOS notation, (2) applying goal modeling heuristics (same as in the first case study), and (3) ensuring the independence of goals in the decomposition.

**Expert Probability Elicitation.** Expert elicitation focused on the leaf goals and the decomposition nodes of the goal models and followed the elicitation protocol described in Section 3.2. For organizing the expert elicitation sessions, we considered three options [19]:

- *Interactive groups*, where experts meet face-to-face with a session moderator, also known as data gatherer.

- *Individual interviews*, where experts are interviewed individually and alone by the data gatherer.

- *Delphi meetings*, where experts do not interact directly and instead provide their answers to the data gatherer in isolation. Afterwards, the data gatherer anonymizes and distributes the judgments to the experts, allowing them to revise their previous answers if necessary.

The experts were briefed about the advantages and disadvantages of each of the above alternatives in relation to proneness to biases (see [19], Chapter 8). The experts decided that interactive groups would be the most suitable method in both case studies. The rationale here was that the experts were applying our approach for the first time; interactive groups allowed them to have open discussions about the approach and the quantities being elicited. They deemed these open discussions to be important for minimizing ambiguity about the quantities.

Since multiple experts were involved in our case studies, we needed a strategy for obtaining a single distribution based on the potentially differing expert opinions about each of the quantities under elicitation. To this end, we considered two options:

- *Behavioral aggregation*, which relies on the experts themselves to arrive at a consensus.

- *Mathematical aggregation*, which uses a mathematical formula for combining expert opinions. Several techniques can be applied for mathematical aggregation, depending on what the expert opinions represent and the way in which a decision maker wishes to interpret the (aggregated) result [19, 1].

For interactive groups and Delphi meetings, behavioral aggregation is most suitable; whereas, for individual interviews, mathematical aggregation is typically applied [19]. In line with our decision to use interactive groups for organizing the elicitation process, we used behavioral aggregation for obtaining a single response for each of the elicited quantities. To ensure convergence, the facilitators explained to the experts that they needed to arrive at a single response for each quantity through persuasion and compromise [19].

A final consideration regarding expert elicitation was how accurate the experts were in providing their estimates. Often, experts who are taking part in elicitation activities for the first time need to undergo special training, aimed at ensuring that the experts are well-calibrated [1], i.e., they are able to estimate their own uncertainty with a high level of accuracy. In our case studies, no calibration training was performed, as all the participating experts had been already involved in several previous projects, where they had provided probability estimates for failures and hazard trigger events.

#### *5.1.5. Results*

Our case studies were structured in the form of interactive workshops with participation from both the researchers and the experts. Over the course of the two case studies, a total of 7 full-day workshops and 2 half-day workshops were held for goal model construction, expert elicitation, and reviewing and adjustment of quantitative assessment outcomes. Four full-day and one half-day workshops were in conjunction with the first case study and the remainder were in conjunction with the second. In the first case study, the high-level goal model specified all known failure mechanisms for fiber ropes in mooring systems. In Figure 2, OB2–OB4 represent a partial list of the known mechanisms. Subsequent goal elaboration focused on a particular aspect of tension failure (OB3), caused by the time-dependent behavior of fiber ropes. The quantitative assessment part of our case study considered the branch rooted at G3, indicated with a dashed line in Figure 2. The elaboration and linking of evidence was done only for the 5 leaf goals that are descendants of G3. In total, the branch rooted at G3 has a total of 8 goals and 11 evidence items supporting the five leaf goals in the branch.

The second case study proceeded similarly to the first, by identifying the main failure mechanisms that could compromise the safety of underwater installation operations. The experts identified three general mechanisms: (1) rope failure, (2) failure of mechanical machinery, and (3) failure of control. Goal elaboration then focused on the mitigation of rope failures in the context of installation. The resulting goal model, not shown due to space constraints, includes 20 goals, obstacles, and assumptions, of which 7 are leaf goals. These leaf goals are supported by a total of 12 evidence items.

Expert elicitation in both case studies focused on the leaf goals and decomposition nodes of the respective goal models. In the first case study,

Table 3: Summary of case studies

| **Case Study** | **Number of goal model elements*** | **Number of elicited quantities** | **Case Study Execution** |
|---|---|---|---|
| Fiber Rope Mooring | 19 | 8 | 4 full-day workshops and one half-day workshop |
| Fiber Rope Installation | 32 | 9 | 3 full-day workshops and one half-day workshop |

**An element is defined as being a goal, obstacle, assumption, or evidence item. The element count considers only elements that are directly related to quantitative assessment in each case study.*

8 quantities were elicited (5 leaf goals and 3 decomposition nodes) and in the second case study 9 quantities (7 leaf goals and 2 decomposition nodes). Table 3 summarizes the main characteristics of our two case studies.

After the elicitation activities were concluded, Monte Carlo simulation and sensitivity analysis were performed. This process was illustrated in Sections 3.3.2 and 3.4. In both case studies, the goal satisfaction and sensitivity results were subsequently presented to the experts for review and possible adjustments. In the first case study, no changes were made by the experts to the elicited quantities. In the second case study, one of the quantities was revised during reviews, as the experts realized that the value given earlier for the quantity in question was for another rope type.

In both case studies, the experts found the results to be intuitive and useful. In particular, they found quantitative analysis as a valuable aid for identifying where expert judgment introduces the most uncertainty and for taking steps to reduce the dependence on expert judgment by provision of more thorough evidence. Initiatives are already underway to develop more detailed evidence collection guidelines where uncertainty was found to be high.

### *5.1.6. Discussion*

Below, we discuss the results of the case studies focusing on answering the research questions that motivated these studies.

**RQ1. Is the approach feasible?** In both case studies, we found goal models and goal decomposition to closely match the reasoning performed by the experts. Although important, this finding was not unexpected given that goal models have long been used for requirements elaboration [31, 29] and safety argumentation [17], and that they are increasingly gaining acceptance for standardization and certification [12].

An important aspect of RQ1 is the feasibility of expert probability elicitation in our approach. For elicitation to be conducted successfully, the experts have to be able to understand the questions, remember the relevant information, identify an internal answer to each question, and map their internal answer to probability estimates. In both case studies, expert elicitation was performed successfully. In particular, the experts in our studies found it natural and common to express their opinions using probabilities.

While we could apply our approach successfully in both case studies, the experience gained from the case studies identified areas where further methodological guidance is required for a more effective application of our approach. In particular, the following observations were made through our interaction with the experts:

- The evidence that supports the leaf goals often relies on parameters that describe the relevant operating conditions for a new technology. In our case studies, these parameters included, among others, ambient temperature, seawater solution, tide height, and level of exposure to sun. Feedback from the experts indicated that they would have liked to see these parameters identified and discussed as part of the goal decomposition process, to ensure that the evidence built to support the leaf goals will take all the relevant parameters into account.

- Over the course of the case studies, the experts expressed the need to develop a glossary to define the terms used in the goal models such as the environmental parameters, the safety margins, and the various evidence items used to support the goals. In response to this need, we developed the feature to define a glossary into the Modus tool (section 4). During the case studies, we realized that a glossary alone, while very useful, was not sufficient. In particular, the relationships between the different concepts in the glossary could not be easily specified. For example, we needed to distinguish concepts such as testing results, analytical models, and historical data, while stating that all these concepts were manifestations of the general concept of evidence. Similarly, the relationship between each evidence type and the operating conditions needed to be captured explicitly. Specifying such relationships is much more effective using a conceptual model (e.g., expressed as a UML class diagram) rather than a flat glossary. The experts stated that it would be very beneficial to have guidelines on how to build a conceptual model for the system under assessment.

- Maintaining maximum independence between goals during goal decomposition is an important factor for the soundness of quantitative assessment. In our case studies, goal independence was achieved through reviews and goal restructuring, but we yet have to systematize the process of verifying goal independence. The experts indicated that they would have found it useful to receive guidance on establishing goal independence.

A mention-worthy point about goal models in the context of our case studies is that goal elaboration concerned non-software requirements. The successful application of KAOS goal models (which have been studied primarily for software systems) in a non-software context suggests a much broader applicability range for goal models, and a promising basis for assessment of mechatronic systems [26], which include a combination of software, electronic, and mechanical parts.

**RQ2. Is the effort involved in the application of our approach acceptable?** The first case study required approximately $1\frac{1}{2}$ person months to finish and the second – approximately 1 person months. The two case studies were performed over a span of 4 months. The effort reported is *exclusively* for the construction of a safety argument over existing evidence, and performing one round of expert elicitation and quantitative assessment. No new evidence item was developed as part of our case studies, nor were any of the existing evidence items modified. Overall, the experts found the level of effort in our approach reasonable. While the experts believed that the construction of a goal model and conducting a fine-grained expert elicitation process could take more time than developing a text-based argument, they agreed that the resulting model and probability distributions were easier to communicate, refine, and reuse than text in natural language.

### *5.2. Survey*

Following the completion of our case studies, we conducted a web-based, anonymous survey to obtain the experts' feedback about their experience with our approach. Below, we discuss the design and outcomes of this survey.

### *5.2.1. Survey Design*

**Quality Attributes (QAs) under investigation.** The goal of the survey was to assess the impact of our approach on TQ along the following eight Quality Attributes (QAs):

- *Learnability:* it refers to how quickly one can understand and use an unfamiliar approach. In TQ, learnability is concerned with the level of effort a new staff member (i.e., someone who possesses the required technical background but has no prior TQ expertise) must invest in order to learn the notations and processes used in TQ, for example in the specification of qualification basis, performing risk assessment, testing, and management of the qualification activities.

- *Productivity:* it refers to how efficiently time and budget resources are used. In TQ, the focuses of interest are the time and budget resources consumed at different stages of the TQ process. An increased productivity in TQ means that the qualifying body, the technology supplier or both manage to get more TQ-related work done while spending the same or less amount of resources.

- *Expressiveness:* it refers to the ability to accommodate, in a precise and complete manner, the range of specification and analysis situations that arise in practice. The higher the expressiveness of an approach, the broader the range of situations that the approach can handle effectively. In TQ, higher expressiveness means being able to specify and reason about a larger set of application domains and systems. For example, a TQ approach that is suited to systems with both mechanical and electronic components is more expressive than an approach that can only handle purely mechanical systems.

- *Traceability:* it refers to the ability to envisage and capture the links between information items that are related in some way. Traceability further covers proper description of the rationale for creation, modification, or use of different information items. In TQ, traceability refers to ability to establish a demonstrable link between the qualification basis, the risks identified, and the qualification evidence.

- *Accuracy:* it refers to how close the outcomes of an activity are to being "correct". In TQ, accuracy refers to the ability to perform a correct assessment of the fitness of a technology for its purpose. The higher the accuracy of a TQ approach, the higher the level of trust in the TQ decisions made as the result of applying the approach.

- *Reusability:* it refers to how easy it is to use the artifacts built in one project, in another related project. In TQ, reusability can be desirable for many artifacts including but not limited to requirements specifications, inspection, testing and assessment results, and environmental parameters of the technology being assessed.

- *Support for collaborative work:* it refers to how easy it is to conduct a project in a collaborative manner, involving a potentially large number of people who may be distributed over different geographical sites or across different organizations. In TQ, the distribution can be due to the presence of multiple stakeholders (suppliers, qualifiers, integrators, operators) and multiple TQ experts who have to develop a consensus during technology verification.

- *Tool support:* it refers to the availability and suitability of tools that support a given approach.

**Questions about experts' backgrounds.** As part of the survey, the respondents were asked about their background in TQ and our proposed approach. The related questions are given in Figure 10. We note that, to the experts, the whole approach and not just the tool was known as Modus; therefore, the question about the time spent on Modus in Figure 10 indicates experience with the whole approach. In our survey instructions, we clarified to the experts that the time spent on Modus covered the following: (1) Attending meetings, workshops, and presentations related to Modus, (2) Self-reading of Modus technical reports, (3) Using the Modus tool, (4) Constructing and reviewing of KAOS goal models, and (5) Contributions to writing technical reports on Modus.

**Questions about QAs.** Respondents were asked two questions about each QA, as shown in Figure 11. In the figure, $QA_i$ denotes a given QA, (e.g., learnability) from the list of eight QAs described earlier. In addition, an open box was available to the respondents to provide a qualitative justification for each of the provided answers.

- Experience with TQ (years or months):
- Time spent on Modus (months or hours) since first acquainted with it:

Figure 10: Survey questions on experts' backgrounds

*Q1.* How important is $QA_i$ for running a successful TQ program?

□ Very Important

□ Moderately Important

□ Unimportant

*Q2.* What impact will Modus have on $QA_i$ if the current TQ practice is complemented with Modus?

□ Modus will significantly improve $QA_i$.

□ Modus will moderately improve $QA_i$.

□ Modus will have negligible or no impact on $QA_i$.

□ Modus will moderately reduce $QA_i$.

□ Modus will significantly reduce $QA_i$.

Figure 11: Survey questions on QAs

### *5.2.2. Survey Results*

**Respondents' backgrounds.** We invited eight experts to respond to the survey. The invitees were the 5 experts involved in the case studies, along with another 3 TQ experts at DNV who had followed the progress of the research but were not involved in the case studies. Six of these eight experts responded to our survey. On average, the respondents had 10 years of experience in TQ, and 80 hours of experience with our proposed approach.

**Importance of QAs.** *Q1* (Figure 11) aims at validating the set of QAs adopted to assess our approach. Figure 12 shows the results obtained from the survey for *Q1*. The columns indicate the QAs and the pattern fills indicate the percentage of specific respondents' answers.

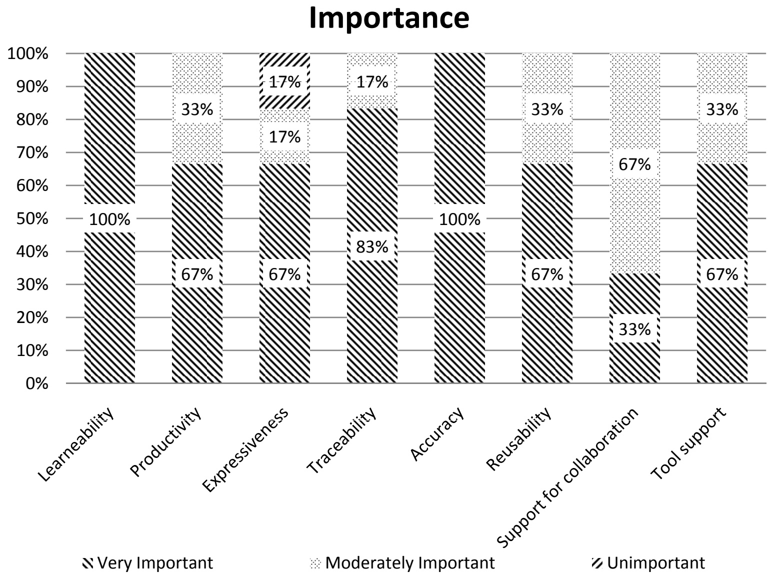


Figure 12: The importance of the QAs under investigation

According to the figure, except for "Support for collaboration", all other QAs were deemed as "Very important" by more than two thirds of the respondents. "Accuracy" and "Learneability" were deemed unanimously as "Very Important". "Unimportant" appeared only once in the responses, selected by a respondent for "Expressiveness". The results therefore suggest that the QAs chosen for the survey are of high relevance to TQ.

**Impact of our approach on QAs.** *Q2* in Figure 11 aims at investigating the impact of the adoption of our approach. Figure 13 shows the results obtained for this question over different QAs. According to the figure, the adoption of our approach was perceived as making improvements to all QAs by at least two thirds of the respondents. The results were unanimous for "Traceability", which was found by all the respondents to be moderately or significantly improved as the result of introducing our approach. A significant negative impact for "Expressiveness" was seen by one of the respondents but no qualitative justification was provided by the respondent.

An interesting but unexpected observation from the survey was about Learnability. We anticipated Learnability to be affected negatively: the adoption of our approach entails learning about goal modeling and expert elicitation in the context of goal models, which we saw as an overhead for learning.

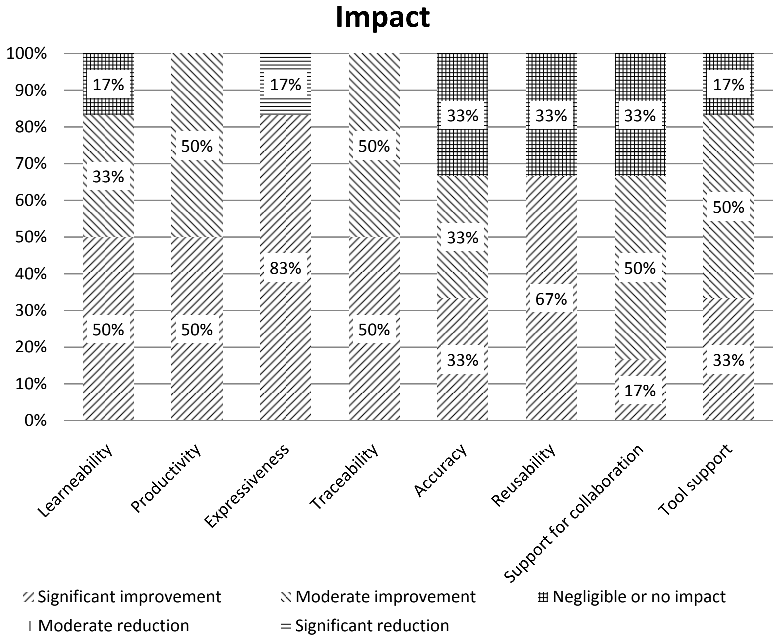


Figure 13: The impact of adopting our approach

To understand the reason why the experts saw learnability improvements, we had a follow-up discussion with them. We found out that the experts saw the additional effort for learning goal modeling and expert elicitation as negligible, when considering the contributions that our approach makes to learnability by more succinctly organizing the existing knowledge about the overall TQ process.

## 6. Discussion

In this section, we discuss some practical considerations and limitations concerning our approach.

**Choice of goal modeling language.** While the theory behind our approach is agnostic to what goal modeling language is being used, the choice of the language is of great practical importance. In particular, from a technology transfer standpoint, it may be advantageous to choose a language that builds on standardized notations and their extensions. This helps to mitigate lock-in to proprietary tools and to ensure that the selected language is going to be supported for a long time. To our knowledge, there is currently no industrial standard for the KAOS language which might be viewed as a limitation for technology transfer. There is work in progress however on a standardized and generic Argumentation Metamodel (ARM) [2]. A semantic mapping from the concepts in KAOS to ARM will allow one to use KAOS with the future ARM-based tools, but developing such a mapping will require further investigation.

**Experts' backgrounds.** The ease at which expert elicitation is performed might depend on the experts' backgrounds. In our case studies, the experts were highly familiar with probabilistic reliability analysis and regularly used such analysis in their day-to-day work. Thus, no training was necessary on probability theory and no calibration was required for expert elicitation. Further, it was easy to convey to the experts how the results of quantitative goal assessment should be interpreted. While our approach draws on standard engineering statistics, we cannot ascertain that the same level of familiarity with probabilistic analysis exists in all domains where our approach may be applicable. Therefore, there may be an additional learning curve associated with our approach which our current case studies do not capture.

**Generalizability.** The probabilistic reasoning model that we use in our work is aligned with the notion of Safety Integrity Levels (SILs) in major safety standards for programmable electronic systems, e.g., IEC 61508 [11]. This standard specifies four SILs (numbered 1–4) for safety functions, with SIL1 being the lowest and SIL4 – the highest. Each SIL is defined as a range for the average probability of failure on demand for low-demand modes of operation, or the probability of a (dangerous) failure per hour for high-demand or continuous modes of operation. While our case studies so far have involved only mechanical components, based on the argument above, we anticipate that the core principles of our work would be applicable to a broader class of systems.

## 7. Related Work

Our work is inspired by and builds on the notion of assurance cases [13], and more specifically safety cases. A safety case is defined as a structured set of evidence-supported arguments to demonstrate that a system is acceptably safe for a given application in a given context [17]. The most adopted framework for safety-case construction is the Goal Structuring Notation (GSN)

[17]. GSN enables analysts to define and decompose goals in a similar manner to KAOS [29] – the goal language we use. Our motivations for choosing KAOS were described earlier in Section 3.1. Further, despite being founded on the same principles, there is a subtle but important conceptual distinction between the notion of "goal" in GSN and that in KAOS. Specifically, GSN is concerned with "argumentation" goals, i.e. claims, whereas KAOS is concerned with "system" goals and obstacles. In GSN, while OR can be used for decomposing and exploring alternative argumentation strategies, it is expected that the final argumentation structure, i.e., the structure that is subject to formal assessment by a third-party, should be OR-free. Conversely, in the context of our work, OR is necessary for refinement of obstacles and also to denote the situation where several alternative measures (goals) are realized by the system to mitigate a particular risk. With this distinction recognized, our approach can be adapted to work with GSN as well.

In addition to GSN, we know of two other goal-based approaches that are targeted specifically at structuring and analyzing assurance cases. These are Trust-IT [4] and Property-Part Diagrams [15, 16]. The argumentation framework in Trust-IT is similar to GSN but its assessment method is quantitative, as opposed to qualitative, which is the case for GSN. We share with Trust-IT the motivation for quantitative assessment of assurance cases, but use a different mechanism for quantification. Specifically, the basis of quantification in Trust-IT is Dempster-Shafer theory of beliefs [27]; whereas we use probability theory. In addition to being in line with TQ current practices, the use of probability theory has two main advantages: (1) the existence of proven guidelines for expert elicitation of probabilities [19, 1]; (2) the flexibility offered by probability theory to conduct advanced analyses such as sensitivity (Section 3.4).

Property-Part Diagrams combine KAOS-like goal models with Problem Frames [14] to formally model the dependencies from critical requirements to environmental assumptions and the behavior of system components. This in turn enables formal reasoning about critical requirements and localizing these requirements to a small subset of system components that contribute to the satisfaction of the requirements. The general principles we apply for requirements decomposition in our work are similar to that in Property-Part Diagrams, but the two approaches differ in the way they reason about requirements satisfaction. Whereas our approach is stochastic, the reasoning performed over Property-Part Diagrams is exact and based on formal logic.

Our approach uses goal propagation for computing a degree of satisfaction for goals. Goal propagation is a topic that that has been studied in the (Software) Requirements Engineering for a long time [8, 10]; however, the focus of the existing literature is on propagation of point-values. A notable exception [9] concurrent to our work uses a combination of simulation and search-based techniques for analyzing tradeoffs in quantitative goal models. Our work applies the same mathematical ideas for simulation, but is targeted at safety and reliability quantification as opposed to tradeoff analysis. In addition, our work includes tailored expert elicitation guidelines and sensitivity analysis facilities which are not within the scope of [9].

Our analysis of goal satisfaction further bears similarity to Fault Tree Analysis (FTA) [6]. Specifically, both fault trees and the goal models in our approach, also expressed as trees, elaborate the relationships between the different elements in the tree using Boolean operators. A complementary notion to fault trees are event trees [6]. In contrast to fault trees, event trees elaborate sequences of events that are linked by conditional probabilities, e.g., event $B$ leads to event $A$ with probability $P(A \mid B)$. Event trees are most suited for capturing continuity in a sequence of conditional events, whereas fault trees are most suitable for specifying different failure scenarios through logical operators [6]. In terms of expressive power, our approach closely resembles fault trees with the difference that in our analysis, we concentrate on goal satisfaction (success scenarios) as opposed to failure scenarios.

## 8. Conclusion

In this article, we presented a tool-supported approach for qualification of new technology. The main novelty of our work lies in seamlessly combining goal modeling, expert elicitation, and probabilistic simulation for quantitatively assessing the satisfaction of a technology's safety and reliability goals. Our software tool unifies the various aspects of our approach into a coherent implementation. We applied our approach in two case studies in the offshore domain and assessed it through a survey involving experts.

Our approach is aimed at providing quantified estimates for the satisfaction of safety and reliability goals based on probabilistic assessment. To be able to trust its quantitative outcomes, our approach needs to be complemented and applied in tandem with *qualitative* measures, both to ensure the precision and adequacy of the goal models built, and to properly reflect on the qualitative insights that experts will have inevitably brought to mind during the probability elicitation process.

In future work, we would like to provide support for quantitative cost and performance comparisons between alternative choices in the technology design phases. The existing literature on trade-off analysis for security [3] and requirements engineering [9] is a promising starting point in this direction. We would further like to develop ways to model and aggregate different types of system decomposition for analysis. System decomposition often encompasses three main viewpoints: the process taken to develop a system, the system's structure, and the system's behavior. Being able to reason about the dependability of a system in a holistic manner requires the integration of these different viewpoints. As a first step, we are investigating the integration of the behavioral and structural viewpoints, particularly based on existing guidelines such as AADL's Error Model Annex [25] and the AltaRica language (`http://altarica.labri.fr/`).

**Acknowledgments.** We are very grateful to the staff at DNV's Technology Qualification Service Line for sharing their valuable knowledge and insights with us, and for their continued help in all aspects of the work reported in this article. We would like to specially thank Dag McGeorge, Vidar Åhjem and Jonas Borg for their commitment and effort in aligning our work with industrial needs, improving the tool support, and conducting the case studies. We gratefully acknowledge funding from the National Research Fund, Luxembourg (FNR/P10/03 - Validation and Verification Laboratory), from DNV under the ModelME! project, and from the Research Council of Norway under Project No. 203461/030.

## References

[1] A. O'Hagan et al. *Uncertain Judgements: Eliciting Experts' Probabilities.* Wiley, 2006.

[2] Argumentation Metamodel (ARM). version 1.0 (2nd FTF) - beta 2. http://www.omg.org/spec/ARM/1.0/Beta2/.

[3] S. Butler. Security attribute evaluation method: a cost-benefit approach. In *ICSE'02*, pages 232–240, 2002.

[4] L. Cyra and J. Górski. Expert assessment of arguments: A method and its experimental evaluation. In *SAFECOMP'08*, pages 291–304, 2008.

[5] J. Devore and N. Farnum. *Applied Statistics for Engineers and Scientists.* Duxbury, 2nd edition, 2004.

[6] C. Ericson II. *Hazard Analysis Techniques for System Safety.* Wiley, 2005.

[7] P. Garvey, editor. *Probability Methods for Cost Uncertainty Analysis.* Marcel Dekker, 2000.

[8] P. Giorgini, J. Mylopoulos, E. Nicchiarelli, and R. Sebastiani. Formal reasoning techniques for goal models. *J. Data Semantics*, 1:1–20, 2003.

[9] W. Heaven and E. Letier. Simulating and optimising design decisions in quantitative goal models. In *RE'11*, 2011.

[10] J. Horkoff and E. Yu. Comparison and evaluation of goal-oriented satisfaction analysis techniques. *Requirements Engineering Journal*, 2011. (in press).

[11] IEC 61508: Functional safety of electrical/electronic/programmable electronic safety-related systems, 2005. Intl. Electrotechnical Commission.

[12] International Maritime Organization (IMO). Goal-based new ship construction standards: Annual working group report. `http://www.imo.org/safety/mainframe.asp?topic_id=1017`, Last accessed 2012.

[13] D. Jackson, M. Thomas, and L. Millett. *Software for Dependable Systems: Sufficient Evidence?* National Academy Press, 2007.

[14] M. Jackson. *Problem Frames: Analysing and Structuring Software Development Problems.* Addison-Wesley, 2001.

[15] E. Kang. A framework for dependability analysis of software systems with trusted bases, 2010. Available at: `http://people.csail.mit.edu/eskang/papers/eunsuk_ms.pdf`.

[16] E. Kang and D. Jackson. Dependability arguments with trusted bases. In *RE*, 2010.

[17] T. Kelly and R. Weaver. The goal structuring notation - a safety argument notation. In *Dependable Systems and Networks 2004 Workshop on Assurance Cases*, 2004.

[18] Managing the safe service life of fiber ropes for mooring. Technical report, DNV, 2009.

[19] M. Meyer and J. Booker. *Eliciting and analyzing expert judgment: a practical guide*. SIAM, 2001.

[20] A. Mosleh. Interaction between model and data in common cause failure analysis. Technical Report B9-13, U. Maryland, 1989.

[21] Position mooring. DNV-OS-E301, Det Norske Veritas (DNV), 2010. Available at: `http://exchange.dnv.com/publishing/codes/download.asp?url=2010-10/os-e301.pdf`.

[22] Qualification procedures for new technology. DNV-RP-A203, Det Norske Veritas (DNV), 2011. Available at: `http://exchange.dnv.com/publishing/Codes/download.asp?url=2011-07/rp-a203.pdf`.

[23] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2005.

[24] M. Sabetzadeh, D. Falessi, L. Briand, S. Di Alesio, D. McGeorge, V. Åhjem, and J. Borg. Combining goal models, expert elicitation, and probabilistic simulation for qualification of new technology. In *HASE*, 2011.

[25] SAE AADL Annex Volume 1: Annex E: Error Model Annex, 2006.

[26] W. Schafer and H. Wehrheim. The challenges of building advanced mechatronic systems. In *FOSE '07*, pages 72–84, 2007.

[27] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton Press, 1976.

[28] Technology qualification management. DNV-OSS-401, Det Norske Veritas (DNV), 2010. Available at: `http://exchange.dnv.com/publishing/Codes/download.asp?url=2010-10/oss-401.pdf`.

[29] A. van Lamsweerde. *Requirements Engineering: From System Goals to UML Models to Software Specifications*. Wiley, 2009.

[30] P. Walley. *Statistical reasoning with imprecise probabilities*. Chapman and Hall, 1991.

[31] E. Yu. Towards modeling and reasoning support for early-phase requirements engineering. In *RE'97*, pages 226–235, 1997.