On Eliciting Contribution Measures in Goal Models

Sotirios Liaskos
School of Information Technology
York University
Toronto, ON, Canada
liaskos@yorku.ca

Rina Jalman
School of Information Technology
York University
Toronto, ON, Canada
rjalman@yorku.ca

Jorge Aranda
Dept. of Computer Science
University of Victoria
Victoria, BC, Canada
jaranda@uvic.ca

Abstract—**Goal models have been found to be useful for supporting the decision making process in the early requirements phase. Through measuring contribution degrees of low-level decisions to the fulfilment of high-level quality goals and combining them with priority statements, it is possible to compare alternative solutions of the requirements problem against each other. But where do contribution measures come from and what is the right way to combine them in order to do such analysis? In this paper we describe how full application of the Analytic Hierarchy Process (AHP) can be used to quantitatively assess contribution relationships in goal models based on stakeholder input and how we can reason about the result in order to make informed decisions. An exploratory experiment shows that the proposed procedure is feasible and offers evidence that the resulting goal model is useful for guiding a decision. It also shows that situation-specific characteristics of the requirements problem at hand may influence stakeholder input in a variety of ways, a phenomenon that may need to be studied further in the context of eliciting such models.**

Keywords-**requirements engineering, goal modelling, prioritization, analytic hierarchy process**

I. Introduction

Goal models have been recognized as a promising approach for modeling and reasoning about alternative solutions during early requirements [1], [2], [3], [4]. Through refinement hierarchies, such models represent alternative ways by which top-level stakeholder goals can be fulfilled. Each of the identified alternatives is assessed subject to quality requirements, through modeling the *contribution* of low-level decisions to the fulfillment of each such quality. Based on the relevant significance of these contributions, solutions that best match stakeholder priorities are identified and pursued. A wealth of approaches that allow such reasoning has being proposed in the literature, employing a variety of measuring scales and inference techniques [5].

However, the problem of eliciting/identifying the contribution measures, that is, the question *where the numbers come from* [6], has not enjoyed equal attention in the research community. How can such contribution assessments be elicited from stakeholders and formalized in concrete measures and what factors can affect such elicitation? How can these measures be aggregated and used for making a decision and why do we believe that that decision truly represents stakeholder input?

In this paper, we explore the use of the Analytic Hierarchy Process (AHP) [7] for eliciting and aggregating quantitative contribution measures within semi-formal goal models, aiming at supporting decision making during early requirements analysis. In particular, we are exploiting the similarity between goal hierarchies which are central in goal models and criteria hierarchies, which is how AHP organizes priority elicitation. Thus, we propose that, for goal models that meet certain structural characteristics, the problem of eliciting contribution levels and making a decision can be seen as an aggregation of a number of individual standard AHP decision problems. Solving each of these problems is effectively a way to assess the contribution measures within the goal model and eventually decide over a solution through a simple reasoning procedure. This way, goal modelers can present a stronger validity argument both for the resulting representation and for the decisions it yields.

To understand different aspects of this synergy in practice, we also conducted an exploratory experiment. We gave the classic meeting scheduling problem in form of a goal model to a group of participants and asked them to envision themselves in given scheduling scenarios. They followed AHP for eliciting contribution measures based on each scenario. We observed the participants for their consistency in their responses within and across scenarios as well as subsequent recognition of their preferences and alternatives of choice. Amongst the findings are that, for a small problem: (a) the process is applicable and soon converges to consistent or near-consistent results, (b) participants generally recognize the result of their preference input and preferred choice, (c) situational characteristics (i.e. the particular decision making scenario) may influence some but not all of the priorities and (d) quantitative representation of contribution degrees does not seem to impair visual reasoning about contributions and preferred decisions compared to qualitative.

The paper is organized as follows. In Section II we provide an overview of goal models and the AHP process and in Section III we show how we apply the latter to assess the contribution measures of the former. In Section IV we present the design and results of our exploratory study. Then, in Section V we discuss related work and conclude in Section VI.

II. Background

A. Early Analysis Using Goal Models

Goal models allow representation and reasoning about how goals of stakeholders relate with and influence each other. Such a model, adapted from [1], can be seen in Figure 1. The model describes goals pertaining to the classic Meeting Scheduling problem, and ways by which such goals can be achieved. The notation makes use of core concepts that can be found, in one form or another, in most dialects of the *i** family (e.g. [8], [9]).

In the model, high-level hard-goals – the ovals – are recursively decomposed into lower-level ones. The decompositions are modeled through two kinds of links, the *AND-decomposition* links and the *OR-decomposition* links. All children of AND-decompositions need to be fulfilled for the parent goal to be considered fulfilled. Respectively, fulfillment of just one child of an OR-decomposition suffices for us to consider its parent fulfilled. Thanks to the existence of OR-decompositions the goal tree implies a great number of alternative ways by which the root-level goal can be satisfied. These are simply solutions of the AND/OR tree.

Soft-goals – the cloud-shaped elements – also represent goals, but ones for which there is no cut-and-dry satisfaction criterion. As such, their satisfaction is assessed on the basis of satisfaction of other goals. This is traditionally expressed with *contribution links*: a positive (respectively, negative) contribution link drawn from a goal to another means that evidence of satisfaction of the former constitutes evidence for the satisfaction (resp. denial) of the latter.

Contribution links allow us to view soft-goal satisfaction as a *criterion* for assessing the satisfaction of other higher-level goals. In the Figure 1, satisfaction of soft-goal *Minimal Effort* is influenced by goals *Minimal Collection Effort* and *Minimal Matching Effort*. Our knowledge of satisfaction of *Minimal Effort* is, thus, assumed to depend exclusively on what we know about the satisfaction of *Minimal Collection Effort* and *Minimal Matching Effort*, and, as such, the latter are the criteria for assessing satisfaction of the former. Thus, soft-goals that vary in their level of specificity, naturally form hierarchical structures in which soft-goals of the lower level serve as criteria for assessing the fulfillment of those at the higher level. At the lowest level, contributions to soft-goals originate from hard-goals. Thus, different solutions of the AND/OR decomposition tree imply a different contribution to the satisfaction of lowest-level soft-goals and, in turn, the entire soft-goal hierarchy. Conversely, different desiderata regarding satisfaction of the soft-goal hierarchy, imply different criteria that need to be met and, in turn, a different goal alternative to be considered.

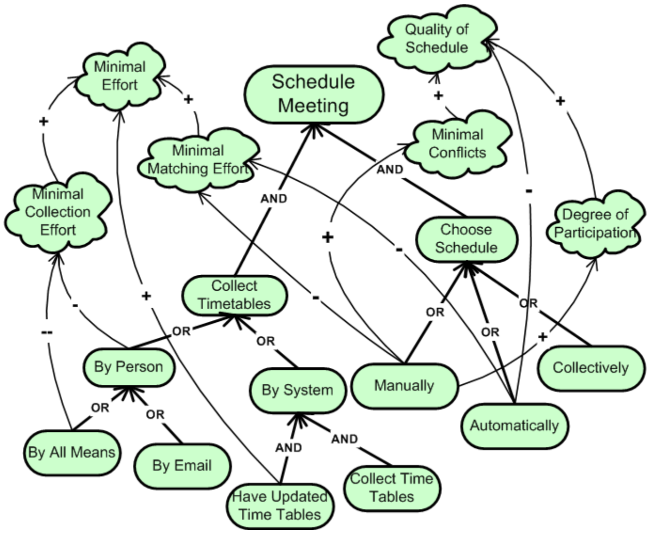

Figure 1. A goal model

But how can contribution measures be elicited? How can they be aggregated when multiple such target the same goal? And how can an informed decision be made based on this? In this paper, we show how we can appeal to the Analytic Hierarchy Process (AHP) to both elicit contribution measures and appropriately aggregate them for the purpose of making decisions within goal models. By doing this we both (a) allow goal models to be the basis for structuring and conducting an early requirements decision problem and (b) supply the representational result (i.e. the resulting goal model with all its elicited weights) with a stronger validity argument. Before we see how this is possible, we take a closer look at AHP in the following subsection.

B. The Analytic Hierarchy Process

The Analytic Hierarch Process (AHP) [7] is a decision support method aiming at quantifying relative priorities for a given set of alternatives based on the subjective judgment of a decision maker. It has been widely used for decision making in many areas like economics, social and management science [10] and in requirements engineering [11]. In AHP complex problems are modeled in a hierarchical structure showing the relationships of the main decision goal, its satisfaction criteria and different levels of subcriteria thereof. More specifically a sequence of five steps is performed which we describe below.

Constructing the Criteria Hierarchy. The first step is the formation of a criteria hierarchy. The top level element of the hierarchy represents the objective that needs to be met, in form of a problem statement – the *decision goal*. For example, if our decision problem is to e.g. purchase a bicycle, the root decision goal would be *Choose Best Bike* (see Figure 2). At the same time a set of *alternatives* that can potentially fulfil the objective are identified. In our bicycle example, three such bicycle models may exist, say, A1, A2, and A3, each with different features. Subsequently, one level under the top goal, the major criteria for assessing fulfillment of the decision goal are defined in broad terms. For example, if it is a bike for long commutes we may be interested in its *comfort* and *durability* as top-level criteria. Each criterion

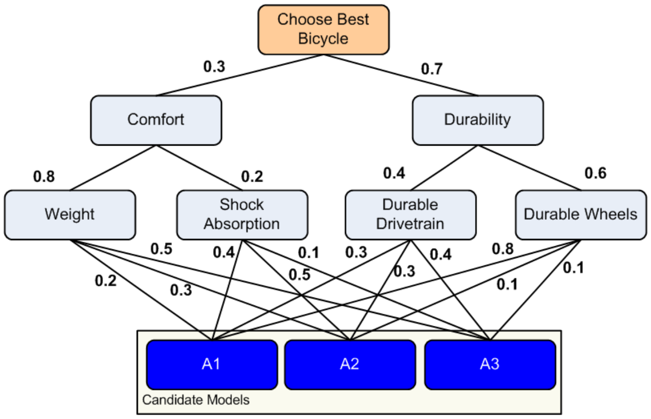

Figure 2. An AHP decision hierarchy

may be broken down to lower-level ones depending on how much detail is needed. In our bicycle selection problem, the comfort criterion involves two sub-criteria: *weight* and *shock absorption* (against e.g. pavement bumps). Once the criteria hierarchy is complete, the alternatives are connected to each of the leaf-level criteria, forming the bottom level of the hierarchy. Thus, the resulting model, such as that of Figure 2 describing our simple bicycle selection problem, has three levels: decision goal, criteria and alternatives.

Pairwise Comparisons and Comparison Matrices. At the next stage, comparisons are performed at each level of the hierarchy, in order for the relative importance of the sibling criteria or alternatives to be assessed. In such comparisons the elements are compared *with respect to their parent element* in the hierarchy. Thus, the top level criteria are compared with each other with respect to their relative importance in achieving the decision goal. In our bicycle example we compare comfort with durability with respect to how important is each in selecting the best bike. Sub-criteria of every level are likewise compared with respect to their importance for fulfilling the parent criterion. Thus, the durability of the drivetrain is compared with the durability of the wheels with respect to how important each subcriterion is for the overall durability of the bike. At the leaf level, alternatives are compared with respect to their fulfillment of each of the lowest level sub-criteria. Overall, for the decision model shown in Figure 2, seven comparisons need to be made: one for comparing top level criteria with respect to the decision goal, two more for comparing the second-level sub-criteria and four for assessing the "goodness" of alternatives with respect to each of the four sub-criteria.

The comparisons are performed in a pairwise fashion. An $n \times n$ matrix is constructed, the *comparison matrix*, where each row and column represents each of the n elements to be compared. Each cell of the matrix hosts the result of the pairwise comparison between the corresponding elements. The decision maker fills each cell with a value expressing: (a) in the case of criteria, the relative importance of one sub-criterion (row) over the other (column) with respect to the parent criterion or (b) in case of alternatives, how much better one alternative is judged to satisfy a given leaf-level criterion than the other. The values are chosen from the set $\{1,3,5,7,9\}$, expressing equal, moderate, strong, very strong, or extreme importance (in case of criteria) or betterness (in case of alternatives) of one element over the other. Thus, back to our bicycle purchase problem, we may say that shock absorption is strongly more important than weight with respect to comfort, hence 5. That number would be different if the parent criterion was different (e.g. performance).

Calculation of Local Weights. In this step, the comparison matrices are transformed into weighted priority profiles amongst the involved items, which we call the *local weights*. Thus, at each level of the hierarchy tree, each of the elements at that level acquires a real number from the interval [0,1] representing its relative importance (for sub-criteria) or relative suitability/"betterness" (for alternatives) of the element compared to its sibling elements and with respect to the parent criterion. Hence local weights represent the contribution share of each element to their parent one. The transformation follows the eigenvector method (EVM) – we omit the details and refer the reader to [7] for details. In Figure 2 such numbers appear as labels on the links that connect alternatives or sub-criteria to higher level criteria.

Aggregation of Local Weights into Global Weights. Once the local weights of elements are obtained at different levels of the hierarchy, they are aggregated to obtain *global weights* of the decision alternatives (elements at the lowest level). To calculate the global weight $global(a)$ for alternative a we use the formula:

$$global(a) = \sum_{c_l \in C_l} (\prod_{c_i \in C_l^{root}} local(c_i) \times local(a, c_l))$$

where C_l is the set of criteria c_l subject to which a has been compared, $local(a, c_l)$ the resulting local weight of each such comparison, C_l^{root} is the set of criteria c_i that are ancestors to the criterion c_l, $local(c_i)$ being their local weights. Thus, in the bicycle example of Figure 2, the global weight of alternative A1 is $(0.3 \times 0.8) \times 0.2 + (0.3 \times 0.2) \times 0.4 + (0.7 \times 0.4) \times 0.3 + (0.7 \times 0.6) \times 0.8$.

Decision. The global weights of the alternatives represent the suitability of each in achieving the decision goal – the higher the more suitable. The result tells us not only which alternative is more important – and we may reasonably want to pursue it as such – but also to what degree.

III. Eliciting the Contribution Structure and Making Decisions

A. The Process

Application of AHP to acquisition and aggregation of contribution degrees in goal models is based on two principles: (a) every OR-decomposition in the goal model constitutes a separate decision problem and (b) the soft-goal hierarchy plays the role of the AHP criteria hierarchy for making each

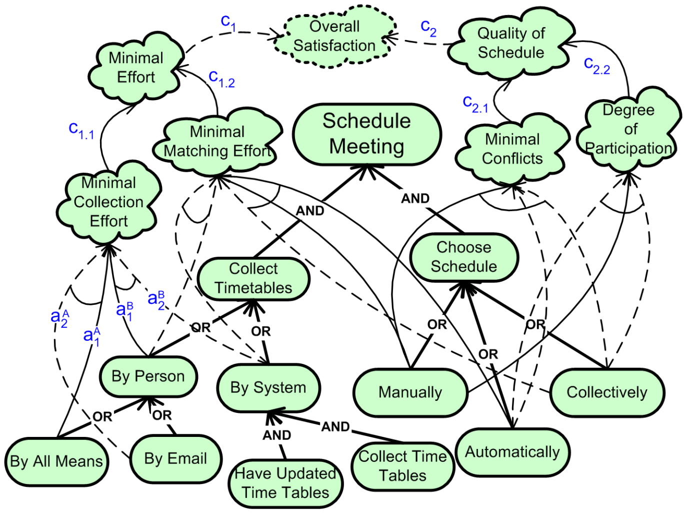

Figure 3. Rearranging and Enriching Contribution Links (added elements appear in dashed lines/borders)

such individual decision. The aggregated result of solving each and every such decision problem is a complete alternative in the goal tree (i.e. a solution of the AND/OR tree). More specifically let us assume that we are given a typical goal tree such as that of Figure 1. The model involves an acyclic (for simplicity) hierarchy of soft-goals, the AND/OR hierarchy of hard-goals and an initial assumption by the modeller on how these are connected through contribution links – though any initial quality and degree of contribution will be ignored in light of the systematic assessment that follows. We then perform the following steps, also presented more concisely in Figure 5.

Rearranging and Enriching links. We first bring the contribution structure to a form that is suitable for the application of AHP. To do so, we firstly ensure that contribution links connecting the hard-goal decomposition with the soft-goals are restricted to ones that connect children of OR-decompositions to leafs of the soft-goal hierarchy. Thus, for links that originate from a goal that is a child of an AND-decomposition, we move the origin of the link to the closest ancestor in the tree that is a child of an OR-decomposition. We remove the link if such ancestor does not exist, since, in this case, the contribution does not serve any purpose for decision or alternatives analysis. Further, once this step is done, contribution links that point to a non-leaf soft-goal are removed and replaced with links that originate from the original hard-goal to each of the leaf-level descendants of that non-leaf soft-goal. The assumption behind this practice is that a contribution to a soft-goal must be explained through contributions to one or more of its lower-level soft-goals (i.e. there are no contribution means that are not being modelled).

As a second step, we ensure that all children of each OR-decomposition contribute to exactly the same soft-goals. To enable this, for every sibling g' of every child g of an OR-node, if g' contributes to a soft-goal s, we draw a contribution from g to s, too, if it does not exist already.

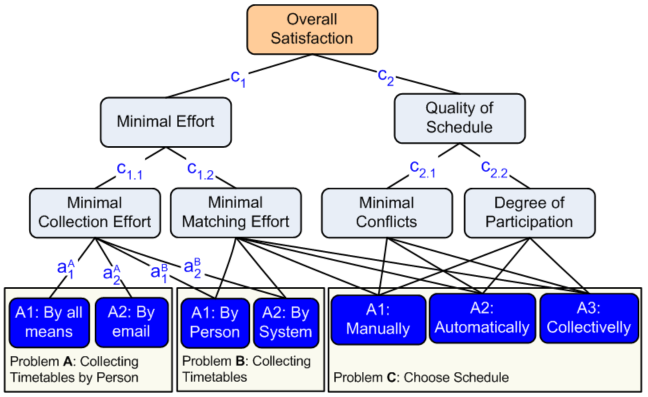

Figure 4. AHP Decision Model for the Meeting Scheduling Problem

Finally, if there are multiple soft-goal hierarchies, a soft-goal "Overall Satisfaction" is introduced to the goal graph and placed as the parent of the roots of each of those hierarchies. This way, soft-goals are all organized in a unique tree structure. The result of applying this step to the goal model of Figure 1 can be seen in Figure 3 – ignore the weight labels for the moment, as the contributions are unlabeled at this stage. In the figure, the links and elements that have been added are drawn as dashed.

From Hard-goals and Soft-goals to Criteria and Alternatives. At this stage, the soft-goal hierarchy is isomorphicaly mapped to an AHP criteria hierarchy. Further, each OR-decomposition is considered as a separate AHP decision problem in which each child of the decomposition is a distinct alternative. All these decomposition problems however are assumed to share the same set of criteria. Figure 4 illustrates how the goal model of Figures 1 and 3 yields three AHP decision problems sharing the same criteria hierarchy.

Acquiring Local Weights. At this step each of the decision problems is solved through AHP pairwise comparisons as described above. Firstly, the local weights in the criteria hierarchy are calculated top-down. Thus, in Figure 4 pairwise comparison between *Minimal Effort* and *Schedule Quality* is performed with respect to *Overall Satisfaction*. This results to local weights c_1 and c_2. Another two comparisons give us the local weights for the lower level criteria $c_{1.1}, c_{1.2}, c_{2.1}$ and $c_{2.2}$. Secondly, each decision problem is completely solved – following an order we will discuss below. For each problem, each of its alternatives are compared subject to their goodness with respect to each of the criteria they have been associated with. The global weights of each alternative are calculated as described earlier and, again, the highest score indicates the alternative of choice.

Mapping Result to Goal Model. The result of AHP analysis has at least two uses with respect to the goal model. Firstly, by mapping the optimal alternative for each

```
/* Preparation */
for every contribution link c: /* $o_c$ and $t_c$ denoting origin and target of c */
  if $o_c$ is a child of AND-node then
    if $o_c$ has ancestors that are children of OR-nodes
    then make c originate from the closest of those ancestors;
    else remove c;
  for every (new) sibling g of $o_c$: draw contribution link from g to $t_c$;
repeat
  for every contribution link c from hard-goal g to soft-goal s:
    if s receives contribution links also from soft-goals $s_1,...,s_n$ then
      {remove c; draw n new contribution links from g to $s_1,...,s_n$;}
until no changes
/* Elicitation */
for every soft-goal s (traversed top to bottom):
  {compare incoming nodes pair-wise subject to s;
   write resulting profiles as local contribution weights;}
for every OR-node g (traversed bottom-up):
  for every soft-goal s to which g's children contribute:
    {compare g's children pair-wise subject to s;
     write resulting profiles as local contribution weights;}
/* Decision */
calculate global weights; pick greatest for each child of OR-node;
```

Figure 5. Eliciting the Contribution Structure using AHP

individual problem to the OR-decomposition from which the problem was generated, we get a solution to the AND/OR tree that is optimal. Secondly, given the one-to-one correspondence of local weights assessed in AHP and contribution links existing in the (enriched) goal model, we can use the former to simply label the latter. Thus, the local weight that expresses the goodness of alternative *By Email* with respect to *Minimal Matching Effort* becomes the label for the contribution link from the hard-goal *By Email* to the soft-goal *Minimal Matching Effort*. Some of the weight labels seen in Figure 4 are transferred to Figure 3 to illustrate how the mapping is done.

B. Discussion

Two aspects of the technique that require some more commenting are (a) the ordering of the decision problems and (b) the reuse of the hierarchy.

Order. Solving each of the AHP problems is ordered based on the hierarchy of the corresponding OR-decompositions in the goal model, starting from the leafs and moving to the top. The reason is that whenever an OR-decomposition is of higher level, each of its alternatives may have descendants that are also OR-decompositions. In such a case, an alternative OR-subgoal may implicitly be a collection of alternatives. In Figure 1, for instance, the alternative to collect timetables *By Person* is in fact two alternatives, one *By Email* and one *By All Means*. Unless we decide on one of those two lower-level alternatives, the comparison between *By Person* and *By System* is problematic. In general, arbitrary such nestings of OR-decompositions may occur. To address this, we consider solving the low level OR-decompositions first and, when we have a higher level OR-decomposition, we mention the optimal solution of each corresponding option in our question to the stakeholder. In our example, if *By Email* is found to be the preferred solution for by-person collection, at the higher level we compare *"By System"* with *"By Person, assuming they do it By Email"*.

Reuse of Criteria Weights. Furthermore, notice that the weights that are elicited for the criteria hierarchy are re-used for identifying the optimal alternative for each problem. To see why this choice, which saves significant effort, is justified, consider that, for example, our preference between *Minimal Effort* and *Quality of Schedule* should not depend on the particular aspect of the meeting scheduling solution we are trying to optimize, but rather on circumstances pertaining to the general scheduling problem.

IV. An Exploratory Study

A. Overview

To assess how the above process applies in practice we performed an exploratory experiment with a number of external participants. The goals of our exploration are four.

Firstly, we want to assess whether the process is at all applicable. AHP has been applied to a wide variety of domains from health care to software specifications. But can we also consider concepts that have traditionally been used in the *i** culture (e.g. goals, soft-goals or goal alternatives), and use them as AHP concepts? To assess this we measure the degree by which participants offer consistent responses both in comparison matrices and in terms of recognising their own preferences in the resulting priority profiles.

Secondly, if we populate the goal model with the numbers we acquire through AHP, is the visual result comprehensible in a way that it can, for instance, allow gauging the optimal decision? To assess this we subject our participants to goal models with predefined contribution measures, and ask them to select the optimal alternative without performing any calculations. We then measure how well their intuition driven by the visual representation matches the AHP-based calculations.

Thirdly, what influences the elicitation process? That is, how do stakeholder responses change under different envisioned situations? To measure this we acquire the same comparison input twice, but each time asking the participants to first imagine themselves in a different situation in which the main goal is to be fulfilled. Then we look for significant differences in the resulting weights.

Finally, what effort does the elicitation process require and would it scale for larger examples? Here we simply time the process of acquiring the comparison input for a small model and attempt a projection for larger models.

To study these we subjected our experimental participants to a series of tasks, as we describe below.

B. Experimental Design

Participants in the study are 10 graduate students of Information Technology, recruited from the first author's class. A minor (3%) part of their mark was offered as inducement. They had experience in goal modeling techniques through academic work they performed for the course; some of them had such knowledge from earlier undergraduate courses. The

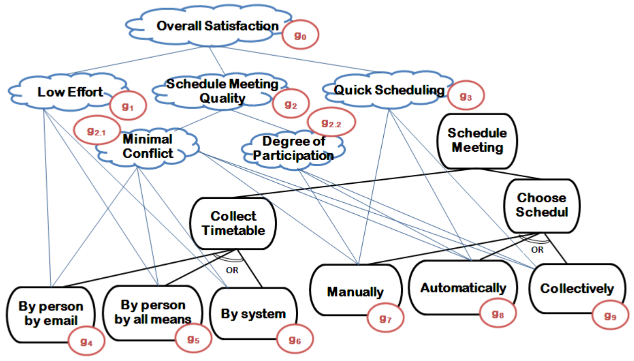

Figure 6. Experimental Goal Model

second author conducted the experiment in her office and supervised it the entire time.

Participants were asked to perform three different tasks, two of which in two different experimental conditions, i.e. in a within-subjects design. The tasks correspond to three different instruments (essentially questionnaires) which we will refer to as Instruments A, B and C. While Instrument A is filled once by everyone in the beginning, Instruments B and C are filled once for each of the two conditions. In each condition participants were given a distinct scenario in which a meeting needs to be scheduled based on the goal model of Figure 6 – a simplified version of the model we saw earlier. The different scenarios allow us to measure the degree by which information about the context of goal fulfillment affects the resulting contribution measures. Thus, in the first scenario (Scenario A), participants envision themselves in a situation where they work on an academic course project with a group of fellow students, they have a very close deadline, and need to meet in order to resolve an unexpected problem. In the second scenario (Scenario B), they are again students in a class who want to organize an end-of-term social gathering with their fellow students for a date that is one month ahead. Participants fill in Instruments B and C for each scenario. The experimental steps are thus as follows.

Step 1 (Instrument A). This instrument aims at assessing the ability of the participants to look at an arbitrary goal diagram with predefined numeric labels and assess what the optimal goal alternative is for that diagram. More specifically, the instrument contains two pairs of diagrams, each pair displaying a different OR-decomposition of the model of Figure 6. For each pair, one model contains numerical values of the interval [0..1]. The values are crafted by the authors to make visual reasoning about the optimal alternative non-trivial, through maximizing conflicting contributions. The other model contains qualitative contribution links "++","+","?", "−" and "−−". The qualitative labels are derived from the quantitative ones in the first model of the pair through fragmenting the continuous space [0..1] into 0.2-step intervals ([0,0.2),[0.2,0.4),. . ., [0.8, 1.0]). Therefore, the qualitative model is roughly a discretization of the quantitative one and, as such, we assumed it to have the same optimal goal alternative. For each model in each quantitative-qualitative pair the participants are asked to choose the optimal alternative based on the given contribution measures, through visual reasoning and without performing any pen-and-paper calculations.

Step 2 (AHP Training). The participants are then given a brief introduction to the AHP method and detailed instructions for filling out the comparison tables. The following Steps 3-6 are performed for each of the two scenarios.

Step 3 (Filling Comparison Matrices). The participants are given the scenario and are asked to imagine themselves as part of the described situation. Based on that situation, they then fill in the AHP comparison matrices. In total, seven (7) comparison matrices are filled for each scenario, as needed for the goal model of Figure 6: two for the soft-goal hierarchy, two for the first OR-decomposition and three for the second OR-decomposition. When needed, and following a recommended AHP practice, the administrator is actively detecting inconsistent responses, through instant calculation of a consistency metric we discus below, and requests participants to revise their preferences into consistent ones (without, of course, dictating the preference per se).

Step 4 (Calculation). For each resulting comparison table the local weights are calculated by following the eigenvalue method. The optimal alternative is calculated based on the aggregation rule we discussed above. Instruments B and C are then given to the participants as follows.

Step 5 (Instrument B). In this instrument we want to measure to what degree participants are able to recognise the local weight profile that best matches their comparison matrix input, according to the eigenvector calculation. The participants are given four (4) comparisons, that are randomly selected from the seven they completed before (Step 3). For each comparison, a number of different options for local weights are provided to the participant. The options are generated as follows: one of them is the one that results from applying the AHP transformations of Step 4 (i.e. the "correct" one) and the rest are all possible permutations of that one. The participant is asked to select the one that best describes their own inputs in the comparison matrix.

Step 6 (Instrument C). In this instrument, we aim at measuring how well the participants' perception of the optimal alternative of the entire goal model matches the AHP-based aggregation of their inputs. Thus, the participants are given four (4) different alternatives for the model of Figure 6 and are asked to select the one that they think best matches their overall priorities that they have been specifying earlier in Step 3. One of the alternatives is the optimal based on participant's input and the AHP aggregation procedure we described, another two are partially optimal (they miss one of the two OR-decompositions – Figure 6) and the fourth is totally incorrect (misses both decompositions).

C. Results

We now take a look at the main results of the experimentation and assess them with respect to our original goals.

1) General Applicability: Our first evaluation goal is to assess whether application of AHP's pairwise comparisons is relevant and doable in goal models. We use two criteria to asses this: firstly, the degree of internal consistency of each individual response, and, secondly, the degree by which the participants are able to recognize their own preferences by looking at the output.

An initial observation is that all participants were able to complete the process without explicating any reservations about the logic or the procedure they followed. Further, the *consistency ratio (CR)* is calculated for each response. The CR tells us to what extend the participant has entered conflicting priorities within a comparison matrix. This is one of the strengths of AHP, as it allows easy detection of otherwise inevitable priority inconsistencies. CR calculation details can be found in e.g. [12]. According to Saaty [7], a value of CR more than 0.1 indicates that judgments should be elicited once again from the participant until she gives more consistent judgments. Yet, participants may refuse to completely lift inconsistencies even if they acknowledge their existence, which happened in our case and is probably why CRs that exceed 0.1 are frequent in practice [12].

Thus, according to the results, the CR values that are greater than 0.1 are about 15% for scenario A and 30% for scenario B, mostly by small amounts. The bar-chart of Figure 7, shows the frequency of CR values that were below 0.1, from 0.1 to 0.2, and above 0.2, for each comparison exercise, for both scenarios. In all cases, more than half of the CRs are below 0.1. The highest CR observed was 0.43.

To assess our second goal, i.e. whether participants are able to recognize their output we look at the results from Instruments B and C. In Instrument B, in which the participants were asked to select the local weights that best match their input in the comparison matrix, in 87.50% and 80% of the responses, for scenario A and B respectively, the participants have correctly selected the local weight profiles that correspond to their pairwise input. In other words, participants were largely able to recognize the numeric priority model that corresponds to their own judgment of the situation as elicited by AHP. In Instrument C, where participants were asked to select the overall goal alternative that they think best matches their priorities, in scenario A 80% of them have selected the alternatives that the AHP aggregation process indicates as preferred. From the remaining, 20% selected one that was partially correct (i.e. got one of the two OR-decompositions right) and none (0%) selected one that was totally incorrect. The corresponding percentages for scenario B are 60% (totally correct) 30% (half-correct) and 10% (totally incorrect).

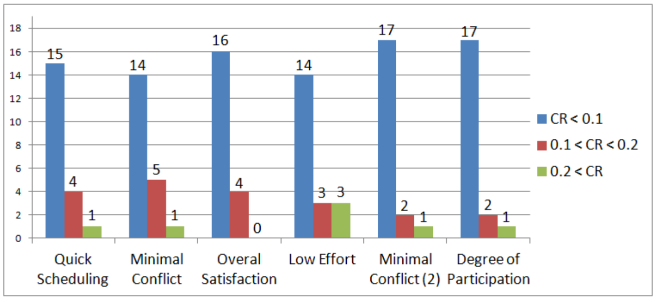

Figure 7. Frequencies of Consistency Ratio Levels

Do these successful responses occur by chance? To investigate this, the binomial test is applied. In Instrument B, the result of the two-tailed binomial test for each scenario was statistically significant ($p < 0.01$ for questions with six options; $p < 0.02$ for questions with two options). The test is statistically significant for Instrument C as well ($p < 0.05$ for four-option questions). Hence, it is highly unlikely that these responses were successful by chance.

Discussion. In analyzing the results we first find the fact that we had consistent or near-consistent inputs as indicative both of relatively reliable data (participants don't answer randomly) and of some basic sanity of the elicitation approach. In Instrument B, the fact that participants are able to recognize *their own* preferences in the profile seems to support the appropriateness of the pairwise comparison method for the purpose we are using it. In other words, what the participants observe in the resulting local priority weights is in agreement with their pairwise input, which seems to validate use of the latter (pairwise comparisons) to produce the former (local weights as measures of contribution share). Furthermore, the ability to recognize not just the local weight but the entire goal alternative (Instrument C) offers evidence on the intuitiveness of AHP-based approach for aggregating local weights: the participants' intuition of what the preferred alternative is coincides to a great degree with the AHP-dictated aggregation of individual local weights, which, in turn, as we saw, are remarkably consistent with the participants' pairwise inputs. In other words, the way the parts are aggregated yields a whole that is consistent with the participants' intuition.

2) The Role of Representation of Weights: Quantitative vs. Qualitative: At the second stage we study the comprehensibility of the weighted measures when placed on the goal model. As we saw, Instrument A offers participants partial goal models with the contribution measures completed by the researchers, in two versions: a numerical and a qualitative. Of the three alternatives they are given in each model, the participants are asked to check the one that is optimal based on the contribution measures. The result is compared with what the AHP-based aggregation process decides as optimal. The participants are also asked to rate their confidence in each answer in a 1–10 scale.

The results show that, in the quantitative models, in 95% of the responses, participants have selected the alternatives correctly with an average confidence score of 8.3/10. In the qualitative models, 75% of the responses are correct and with 7.4/10 average confidence score. To investigate whether the participants are just guessing, the binomial test was again applied: the p-values are < 0.01 for both cases, which are both statistically significant for $\alpha = 0.05$.

Discussion. In this exercise we focus on one of the possible *uses* of the goal model, namely its ability to support a decision. In this context, we investigate the correspondence between, on one hand, the guidance that the visual aspect of the goal model gives to the user regarding how contributions are aggregated and, on the other hand, the mathematical result of the AHP-based aggregation approach. The result suggests a strong connection between the two for both kinds of labels. This allows us to hypothesise that, for a simple model, if we assume validity in the AHP decision making approach, then if its result is placed in the goal model it may also allow visual reasoning over the resulting graphical representation. Another important observation is that the result does not offer evidence that quantitative labelling impairs the effectiveness of the visual reasoning or even the confidence of the respondents. In other words, our observations suggest that adding weights derived from an AHP exercise into a goal model, in either the original numerical or in a converted qualitative form, may aid visual reasoning equally well.

3) The Influence of the Scenario: As a final step, we investigate the role of the scenarios to participant input. One should expect that goals are more or less important from each other depending on the situation in which the goal model is used. But is the participants' input in fact influenced by the scenario? If yes, which part of their input is influenced and which part is not? Is the influence justified or is it a result of an unwanted cognitive bias? To examine this we used two measures. Firstly, we devised a distance measure between comparison matrices. The measure simply calculates the distance between each of the individual inputs in the comparison matrix and sums them up. Thus, for a particular comparison, the more the influence of the scenario the more the distance between the two corresponding inputs. Figure 8 shows a box-plot in which the distances between responses are summarized per goal/comparison for all participants. It is notable that the comparison corresponding to the highest level goal appears to be more dependent on the scenario than the other ones.

We then moved on to assess the statistical significance of the observed difference. Recall that each comparison results in a profile of two or three local weights (depending on the number of items under comparison). For each such individual weight, we tested for significant differences with respect to the scenario change. To avoid Normal distribution assumptions, we used the Wilcoxon test. Twenty such two-tailed tests are performed: out of the seven (7) total comparisons, six (6) are of three (3) choices and one (1) is of two (2) choices. In Table I the Wilcoxon value for each test can be seen – values that are below our significance threshold $\alpha = 0.05$ can be considered significant.

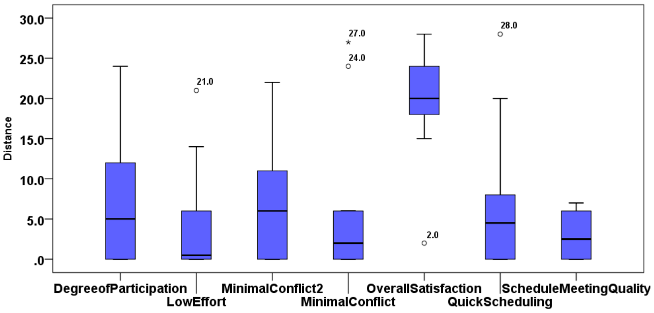

Figure 8. Similarity between comparison matrices across scenarios

Discussion. Both the descriptive analysis and the Wilcoxon test give us a strong impression that some contribution measures in goal models depend on the perceived situation in which goals need to be fulfilled. What is interesting is that not all such contributions are volatile, although they all have the same status in the goal modeling language. In our result, we found that the top level goal is influenced by the scenario the most, as it is more preferential and descriptive of a strategic priority – compared to low-level ones that describe relationships of a more factual nature. We resist however proposing a connection between degree of situation dependency and preferential/subjective nature of a comparison at this point.

Nevertheless, our observation from interacting with the participants, shows another issue pertaining to elicitation: some participants seem to be influenced by the scenario even for aspects where this should not have been the case. For example, when asked to compare timetable collection means with respect to *Low Effort* it became obvious that in scenario A (which involves urgent scheduling), the due date urgency influenced some participants to pick the alternative which allows scheduling as soon as possible, ignoring that the comparison is supposed to be with respect to effort level. This seems to suggest that analysts should be aware that cognitive biases may accompany weight elicitation: participants focus on a piece of information (scenario in our case) that influences their decisions in ways that it should not. In the particular case, the element with respect to which a comparison is supposed to be made (*Low Effort*) is replaced by dominant characteristics of the situation (e.g. urgency). When such biases were detected in the experiment, the intervention of the administrator was often required to carefully (i.e. without dictating a response) remind the participants that the comparison is performed with respect to a very specific criterion. Hence, although we believe that the end-result is not dominated by such biases, the phenomenon is present and needs to be studied more.

4) Effort: AHP is known to pose a scalability issue due to the large number of pairwise comparisons and the

Table I
WILCOXON TESTS PER ELEMENT PER COMPARISON

g_0			$g_{2.1}$			$g_{2.2}$		
$g_1^{(*)}$	**$g_2^{(*)}$**	**$g_3^{(*)}$**	g_7	g_8	g_9	g_7	g_8	g_9
<.01	**<.01**	**<.01**	.89	.5	.89	.46	.92	.67

g_3			$g_{2.1}$			g_1			g_2	
g_7	g_8	g_9	g_4	g_5	g_6	g_4	g_5	g_6	$g_{1.2}$	$g_{2.2}$
.89	.89	.92	.6	.09	.4	1.0	.68	.89	.6	.6

associated effort it may require. From what point does scalability become an issue in AHP's application to goal models? In our application, we found that the process of filling up six (6) 3×3 comparison matrices and one (1) 2×2 takes a maximum of 15 minutes, i.e. about 2 minutes per matrix. The time includes discussing and resolving possible inconsistencies with the participant. Assuming that the effort increases linearly with the number of comparisons, a goal model with, say, one hundred (100) OR-decompositions of three (3) children – which is a realistic average number of children – each decomposition having, say, four (4) different comparison criteria, implies 100×4 comparisons which amounts to slightly less than 15 staff hours, excluding the effort to identify the soft-goal contribution weights – soft-goal graphs are generally much simpler. While it looks like significant effort compared to an ad-hoc approach or no approach at all, for a project large enough to warrant a goal model of hundreds of nodes, it might be actually advisable to invest a few staff-days for increasing validity confidence of the contribution structure and the corresponding decision. However, empirical work seems to be needed in order to validate the linearity assumption behind the above analysis and understand the influence of factors such as e.g. the presence of extended nestings of OR-nodes.

D. Discussion and Validity Threats

We now turn our focus on the general validity of the experiment. The fronts in which *external validity* can be challenged are several. Firstly, we are using only one domain (meeting scheduler) and a particular goal model. Would different goal models for the same domain provide different findings? How about different domains? Further, graduate students of Information Technology seem to be a population that can potentially represent analysts, but cannot possibly represent arbitrary stakeholders, who have a different intuition of numbers and box-and-line models. Furthermore, the small number of comparison tables used in the experiment and the small size of each (three by three maximum), may prevent a straightforward generalization to situations where a larger set of (larger) comparison tables are needed – both in terms of effort (as we saw above) and in terms of quality of input. However, we also believe that the idea of dealing with each OR-decomposition as a separate problem makes size less of a concern in terms of generalization.

Standard measures were adopted to address *internal validity* threats. Thus, as we follow a within-subjects design, counterbalancing was applied to eliminate ordering effect. The comparison tables are given with different, random orders to each participant to eliminate biases in that regard too. Sessions are designed to not exceed 1.5 hours in duration and breaks are taken to avoid fatigue effects. Furthermore, despite our recruitment method (volunteering) we see the threat of self-selection bias irrelevant in our case. Nevertheless, we consider 10 participants to be a small sample despite the statistical significance that emerges by having each complete many exercises. As such, a larger sample size would offer us more confidence in future experimentation.

V. RELATED WORK

The Analytic Hierarchy Process is a well-established multi-criteria decision support method. Several papers have compiled the AHP success stories in very different fields (e.g. [10], [13]). In the literature, there is also considerable commentary describing AHP as a broadly accepted method, based on firm theoretical foundation and, for many, being the most reliable approach to prioritization [14], [15]. In Requirements Engineering the method has been extensively studied as a tool for prioritizing specifications. Karlsson et al. [16] experimentally evaluated six different methods of requirements prioritization and found that AHP is the most promising technique in terms of providing trustworthy results, being fault tolerant, and including consistency measures – Karlsson and Ryan describe the application of AHP in requirements prioritization elsewhere [12]. Nevertheless, we could not find work in Requirements Engineering that makes full use of the hierarchical style of AHP criteria and applies it to goal hierarchies.

In goal modeling there is substantial work on reasoning about goal satisfaction and propagation thereof ([17], [18], [4] – [5] for a survey). Such techniques typically model satisfaction or denial of soft-goals using qualitative or quantitative labels. Contribution operators of various types model how satisfaction of one goal contributes to that of others and how groups of such, potentially conflicting, contributions are aggregated. For example, van Lamsweerde employs score matrices in a manner very similar to the one we consider here [2]. Despite the rigorous theoretical foundation and tool support these proposals offer, they leave the elicitation problem – how we elicit the contribution labels, weights and score values and why we aggregate multiple and conflicting such the way we do – out of their scope.

To address the problem of number elicitation, the literature seems to pursue two directions. Letier and van Lamsweerde [6], propose a probabilistic interpretation of numbers in order to reduce the problem to one of assessing probabilities. This approach, however, cannot be assumed to be applicable to all domains and problems, it may sometimes require too detailed analysis for early requirements and, most importantly, it may not accurately capture stakeholder attitudes, preferences and likings. The latter can perhaps be addressed though the introduction of utility measures. We however could not find an extension to that line of work that makes effective use of utility functions/values – let alone address the problem of elicitation thereof. The second approach to eliciting contribution measures is provided by Horkoff and Yu [19], who propose an interactive conflict resolution technique to address the problem of aggregating influences. The approach seems to be geared towards scalable exploration and comprehension of the model. In comparison, our approach makes fewer presumptions about stakeholder attitudes, requiring the stakeholder to offer their input for every aggregation problem in the goal model using an established elicitation and local weight aggregation procedure. We find the possibility of empirically comparing and combining the two methods very promising.

Finally, the problem of the effect of the visual notation in comprehending goal structures has been studied by Moody et al. [20] through appeal to current theories of perception and cognition. We believe this is a crucial line of investigation which, nevertheless, needs insights from the empirical front in order to shed light on how people understand and use goal diagrams.

VI. Conclusions

We presented an approach for applying the analytic hierarchy process in order to elicit contribution measures in goal models. Soft-goal hierarchies of the goal models are treated as AHP criteria hierarchies and each OR-decomposition of hard-goals is treated as a separate AHP decision problem. The results are plugged back in the goal model, where the optimal goal alternative can be found following the AHP weight aggregation approach. An exploratory experiment offers evidence that AHP's pairwise elicitation is applicable to goals and that both the numeric results and their AHP-based aggregation approach are comprehended by participants.

For the future we wish to tackle certain technical challenges, such as the presence of directed cycles or optional goals in the goal model. A more fundamental issue is that of the *semantics* of numeric values as shares of contribution and their comparison to the existing tradition of contributions as absolute measures of influence. For example, our application reveals that some contribution aggregation cases can be seen as preferential – thus intentional and highly subjective – rather than representations (in any precision scale and assessment method) of factual phenomena of the world. This seems to necessitate a clarification from a conceptual modeling point of view, which we find worth examining.

References

[1] J. Mylopoulos, L. Chung, S. Liao, H. Wang, and E. Yu, "Exploring alternatives during requirements analysis," *IEEE Software*, vol. 18, no. 1, pp. 92–96, 2001.

[2] A. van Lamsweerde, "Reasoning about alternative requirements options," in *Conceptual Modeling: Foundations and Applications*, LNCS, vol. 5600, 2009, pp. 380–397.

[3] P. Giorgini, J. Mylopoulos, E. Nicchiarelli, and R. Sebastiani, "Reasoning with goal models," in *Proceedings of the 21st International Conference on Conceptual Modeling (ER'02)*, Tampere, Finland, 2002, pp. 167–181.

[4] S. Liaskos, S. McIlraith, S. Sohrabi, and J. Mylopoulos, "Representing and reasoning about preferences in requirements engineering," *Requirements Engineering Journal (REJ)*, vol. 16, no. 3, pp. 227–249, 2011.

[5] J. Horkoff and E. Yu, "Comparison and evaluation of goal-oriented satisfaction analysis techniques," *Requirements Engineering (REJ)*, pp. 1–24, 2011.

[6] E. Letier and A. van Lamsweerde, "Reasoning about partial goal satisfaction for requirements and design engineering," in *Proceedings of the 12th International Symposium on the Foundation of Software Engineering (FSE-04)*. Newport Beach, CA, November 2004, pp. 53–62.

[7] T. L. Saaty, *The Analytic Hierarchy Process*. McGraw-Hill International, New York, 1980.

[8] E. S. K. Yu, "Towards modelling and reasoning support for early-phase requirements engineering," in *Proceedings of the 3rd IEEE Int. Symposium on Requirements Engineering (RE'97)*, Annapolis, MD, January 1997, pp. 226–235.

[9] D. Amyot and G. Mussbacher, "User requirements notation: The first ten years, the next ten years," *Journal of Software (JSW)*, vol. 6, no. 5, pp. 747–768, 2011.

[10] O. S. Vaidya and S. Kumar, "Analytic hierarchy process: An overview of applications," *European Journal of Operational Research*, vol. 169, no. 1, pp. 1–29, 2006.

[11] J. Karlsson, "Software requirements prioritizing," in *Proceedings of the 2nd IEEE International Conference on Requirements Engineering (ICRE1996)*, Colorado, USA, 1996, pp. 110–116.

[12] J. Karlsson and K. Ryan, "A cost-value approach for prioritizing requirements," *IEEE Software*, vol. 14, no. 5, 1997.

[13] W. Ho, "Integrated analytic hierarchy process and its applications – a literature review," *European Journal Of Operational Research*, vol. 186, no. 1, pp. 211–228, 2008.

[14] M. Bernasconi, C. Choirat, and R. Seri, "The analytic hierarchy process and the theory of measurement," *Management Science*, vol. 56, no. 4, pp. 699–711, 2010.

[15] E. Triantaphyllou and S.H. Mann, "Using the analytic hierarchy process for decision making in engineering applications: some challenges," *International Journal of Industrial Engineering: Applications and Practice*, vol. 2, no. 1, 1995.

[16] J. Karlsson, C. Wohlin, and B. Regnell, "An evaluation of methods for prioritizing software requirements," *Information & Software Technology*, vol. 39, no. 14-15, 1998.

[17] D. Amyot, S. Ghanavati, J. Horkoff, G. Mussbacher, L. Peyton, and E. Yu, "Evaluating goal models within the goal-oriented requirement language," *International Journal of Intelligent Systems*, vol. 25, no. 8, pp. 841–877, 2010.

[18] P. Giorgini, J. Mylopoulos, and R. Sebastiani, "Goal-oriented requirements analysis and reasoning in the Tropos methodology," *Engineering Applications of Artificial Intelligence*, vol. 18, no. 2, pp. 159–171, Mar. 2005.

[19] J. Horkoff and E. Yu, "Finding solutions in goal models: an interactive backward reasoning approach," in *Proceedings of the 29th International Conference on Conceptual modeling (ER'10)*, Vancouver, BC, 2010, pp. 59–75.

[20] D. L. Moody, P. Heymans, and R. Matulevičius, "Visual syntax does matter: improving the cognitive effectiveness of the i* visual notation," *Requirements Engineering Journal (REJ)*, vol. 15, no. 2, pp. 141–175, 2010.