Modeling and Reasoning with Changing Intentions: An Experiment

Alicia M. Grubb, Marsha Chechik
Department of Computer Science
University of Toronto, Toronto, Canada
{amgrubb, chechik}@cs.toronto.edu

***Abstract*—Existing modeling approaches in requirements engineering assume that stakeholder goals are static: once set, they remain the same throughout the lifecycle of the project. Of course, such goals, like anything else, may change over time. In earlier work, we introduced *Evolving Intentions*: an approach that allows stakeholders to specify how evaluations of goal model elements change over time. *Simulation over Evolving Intentions* enables stakeholders to ask a variety of 'what if' questions, and evaluate possible evolutions of a goal model. *GrowingLeaf* is a web-based tool that implements both the modeling and analysis components of this approach. In this paper, we investigate the effectiveness and usability of Evolving Intentions, Simulation over Evolving Intentions, and GrowingLeaf. We report on a between-subjects experiment we conducted with fifteen graduate students familiar with requirements engineering. Using qualitative, quantitative, and timing data, we show that Evolving Intentions were intuitive, that Simulation over Evolving Intentions increased the subjects' understanding and produced meaningful results, and that GrowingLeaf was found to be effective and usable.**

I. Introduction

Goal modeling has long been used in the literature to model and reason about system requirements, constraints within the domain and environment, and stakeholder goals [1]–[4]. iStar Strategic Rational (SR) Diagrams [1] show dependency relationships between actors and the *intentions* (i.e., goals, tasks, resources, and soft-goals) that rationalize the dependencies. Models consist of actors and their intentions as well as intentions between actors. *Root* intentions are decomposed into alternatives and components until individual evaluable intentions (called *leaf* intentions) are reached. Using a technique called *forward analysis* [5], a modeler assigns evaluation labels (i.e., Fully Satisfied, Partially Satisfied, Partially Denied, and Fully Denied) to each leaf intention and then propagates the evaluation labels to root intentions. Forward analysis gives stakeholders the ability to ask 'what if' questions and find scenarios where desired goals can be achieved.

In earlier work [6], [7], we introduced *Evolving Intentions* (EIs), an approach that includes a language for specifying how evaluations of intentions change over specified time intervals. EIs include a set of pre-defined functions, and allow stakeholders to create new ones (including repeating functions). Once a model has EIs, stakeholders can use *Simulation over Evolving Intentions* (EI-Sim) to explore feasibility of particular scenarios (e.g., will a leaf task eventually satisfy a root goal), and ask a variety of 'what if' questions, using strategies described in [7] that extend forward and backward analysis over time. This analysis is realized in the tool *GrowingLeaf* (see a screenshot in Fig. 1a), described in [8].

In this paper, we investigate the *effectiveness and usability* of EIs, EI-Sim, and GrowingLeaf. We report on an ethics-reviewed between-subjects experiment conducted with fifteen graduate students familiar with requirements engineering (RE). Our experiment aims to answer three research questions:

(RQ1) How do EIs affect modelers' ability to capture model elements that change over time? **(RQ2)** How does EI-Sim affect modelers' understanding and ability to reason about a goal model with time? **(RQ3)** How do modelers evaluate GrowingLeaf after completing modeling and analysis tasks?

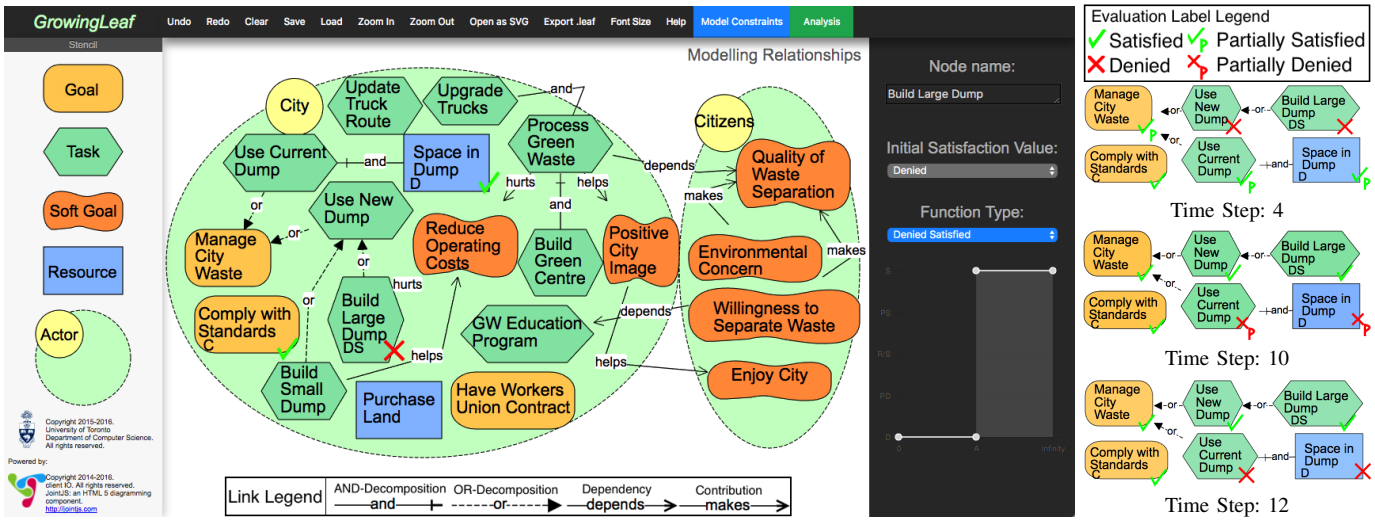
From this experiment, we conclude that the representation of EIs is suitable to the task of identifying and representing intentions over time, and EI-Sim improves the subjects' ability to reason about goal models over time. The subjects also found GrowingLeaf to be usable and effective.

The remainder of this paper is organized as follows. Sect. II introduces EIs and the relevant goal modeling background. Sect. III describes the methodology, materials, and procedure for the experiment. Sect. IV reports the results of the experiment. Sect. V discusses the validity and impact of the study. Sect. VI connects our approach to related work. We conclude in Sect. VII.

II. Background

Running Example (Waste Management). We introduce the relevant modeling background using the Waste Management model from [7], which considers a city evaluating its waste management infrastructure (see Fig. 1a for the model). The current city dump (Space in Dump) has not reached capacity, and the City is considering investing in building a new dump as well as a recycling and composting facility. The City wants to satisfy Manage City Waste, Reduce Operating Costs, and for their citizens to Enjoy City (i.e., their *root* goals).

Goal Modeling. iStar (or i*) SR models consist of actors (i.e., City and Citizens) and their intentions (i.e., goals, tasks, resources, and soft-goals). Intentions can be evaluated using the qualitative evaluation labels *Fully Satisfied* (*FS*), *Partially Satisfied* (*PS*), *Partially Denied* (*PD*), *Fully Denied* (*FD*), *Conflict* (*CF*), *Unknown* (*U*), and *None* (*N*) in the absence of other labels. Intention evaluations are propagated through intention relationships (i.e., links). Decomposition links break



(a) GrowingLeaf modeling view with Waste Management model, used in RQ2, on the centre canvas. Functions shown are: *Constant* (*C*), *Decrease* (*D*), and *Denied-Satisfied* (*DS*).

(b) EI-Sim results at time steps 4, 10 and 12.

Fig. 1: The GrowingLeaf tool modeling view with Waste Management on the centre canvas, and a EI-Sim result.

down intentions into either sub-components requiring all components with *AND-Decomposition*, or sub-options requiring one component with *OR-Decomposition*. When an intention within one actor depends on an intention of a different actor, Depends links connect the intentions through a Dependum[1]. Contribution links indicate influence over the evaluation of soft-goals. The four contribution types are *Makes* (*Breaks*) that propagates sufficient evidence in support for (against) a soft-goal, and *Helps* (*Hurts*) that propagates insufficient evidence in support for (against) a soft-goal.

Forward Analysis. There are a number of techniques for analyzing goal models, summarized in [9]. In this study, we use forward analysis, where leaf intentions give evaluation labels to root intentions via intermediate intentions using the links in the model. Leaf intentions are defined as intentions with no incoming decomposition or contribution links, and no outgoing dependency links. For example, in Fig. 1a, Build Large Dump is a leaf intention and would propagate its value to Reduce Operating Cost and Manage City Waste, but Willingness to Separate Waste would not be a leaf intention because it depends on the GW Education Program.

Evolving Intentions (EIs). We describe how the evaluation of intentions changes over time using a set of functions. *Atomic Functions*[2] define how the evaluation of an intention changes between two consecutive time points. The value can become more true (i.e., *Increase* (*I*)), become more false (i.e., *Decrease* (*D*)), remain *Constant* (*C*), or change randomly, displaying a *Stochastic* (*R*) pattern. For example, if Comply with Standards is *C*, it means that the system always complies. If Space in Dump is *D*, it means that there is less available space over time, and the evaluation label becomes more denied. The *General Compound Function* (called *User-Defined* (*UD*) function) describes any step-wise function specified by modelers as an ordered list of atomic functions. *UD* functions can contain a repeating segment to describe an oscillating behaviour. For example, Build Large Dump is *C* with the value *FD* and then *C* with the value *FS*. Using the General Compound Function, we created a set of *Common Compound Functions* to enable stakeholders to select common functions. For example, the pattern for Build Large Dump discussed above is common, and we define it as *Denied-Satisfied* (*DS*). The opposite, where the evaluation of an intention is *FS* and then *FD*, is defined as *Satisfied-Denied* (*SD*). The *Monotonic Positive* (*MP*) function is used when the evaluation value of an intention increases in satisfaction until it reaches its maximum value and then remains constant at that value. See [7] for a review of EIs.

Simulation over Evolving Intentions (EI-Sim). Once modelers have specified EIs, they can use *Simulation over Evolving Intentions* (EI-Sim). This simulation creates a random possible evolution of the model given initial evaluation labels and EIs in the model. At each time point, the EI-Sim algorithm considers the EI function for each leaf intention and the previous (or initial) evaluation of that intention. Once evaluations of leaf intentions are established, the algorithm propagates values to the root intentions using forward analysis, as described in [7]. The simulation is run for a predefined number of time points. For example, the City wants to know if it can satisfy Manage City Waste over time with the EI function assignments given in the previous paragraph and Fig. 1a. EI-Sim returns one possible evolution of the model (see screenshots of the model fragments in Fig. 1b) where at time point 4 Space in Dump has the value *PS*, and at time point 10 Build Large Dump has transitioned to *FS* and Space in Dump has the value *PD*. Finally, at time point 12, Space in Dump has the value *FD* while Comply with Standards remains *FS* and Manage City Waste remains at least *PS*.

[1]Dependums have been removed for brevity from some figures shown here but appeared in the models shown to the subjects.

[2]Called Elementary Functions in [7].

III. Methodology

The goal of our study is to answer questions RQ1 - RQ3 introduced in Sect. I. For each question, we describe the experimental design and materials required.

The experiment design was developed and tested iteratively with research group members and received peer feedback at RE'16. This experiment was approved by Research Ethics at the University of Toronto. See supplemental information[3] for the study protocol, videos, the list of questions, and handouts.

A. Experiment Design

To investigate our research questions, we need to evaluate subject cognition. Bloom's taxonomy [10] (cognitive domain) defines six dimensions or levels of learning: remember, understand, and apply (ordered) followed by analyze, evaluate, create (parallel). We structure our questions to elicit cognition from a variety of levels (underlined) in Bloom's taxonomy aiming for coverage across the levels.

RQ1. In asking RQ1, we wanted to understand the effectiveness of EIs (introduced in Sect. II). We defined *Stochastically Evolving Intentions* (SEIs)[3] as a control for comparison. SEIs change the evaluations of model intentions stochastically. From this, we generated four sub-questions: (i) How do subjects answer understanding questions about EI and SEI functions? (ii) How do EIs and SEIs affect subjects' evaluation of changing intentions in the model? (iii) What representation would subjects create to indicate intentions with changing evaluations? (iv) How does priming with EIs and SEIs affect subjects' evaluation of the model using forward analysis?

RQ2. In order to answer RQ2, we wanted to compare EI-Sim[4], introduced in Sect. II, with *Simulation over Stochastically Evolving Intentions* (SEI-Sim)[5] and *Repeated Forward Analysis* (Rep-FA)[3]. SEI-Sim generates a new random value, independent of the previous value, for each leaf intention at each time step, and then performs forward analysis. Rep-FA does not actually generate a simulation; instead, we asked the subjects to repeatedly use forward analysis creating their own user-generated simulation manually and updating values at each iteration. Our null hypothesis was that there is no additional benefit to using EI-Sim over SEI-Sim or Rep-FA, in the subjects' ability to reason about the evaluation of intentions over time. We broke down this question by looking at how in the context of each analysis technique, the subjects would (i) identify (understand) and analyze alternative decisions, (ii) analyze the model given a changing intention, and (iii) evaluate the ordering for the completion of two tasks.

RQ3. In order to answer RQ3, we asked the subjects to evaluate GrowingLeaf and provide constructive feedback. We compared the evaluations for all study groups to see if there was a particular part of the tool that was better.

RQ0. In order to investigate RQ1 and RQ2, we had to teach the subjects EIs and/or SEIs resulting in a learning effect. Thus, we evaluated all of our research questions between-subjects instead of within-subjects.

(RQ0) Do modelers perform similarly on basic cognition tests, given a consistent training protocol?

In order to answer RQ0, the subjects were tested on their ability to answer a series of remembering and understanding questions about iStar goal modeling elements and forward propagation over links, as well as the definition of leaf and root nodes. Our null hypothesis was that the subject groups (see Sect. III-C) performed equally well on the questions. We also tested the subjects' ability to use the tool and apply forward analysis as a base line for completing the rest of the study.

B. Materials: Models, Tools, and Videos

We now introduce the materials we used in the study in order to ensure a consistent experience for each subject.

Models. We used three models in this study. For RQ0, we used the *Trusted Computing* model [11] (see Fig. 2a), which shows the relationships between a PC Product Provider, a PC User, and a Data Pirate surrounding the legal or pirated version of PC products. The *Network Administrator* model (see Fig. 2b) was created for RQ1 of this study. Its single actor, the Network Admin is considering which tasks should be completed and in what order, with the goal to Improve Network Infrastructure and Increase Capacity. The third model, used in RQ2, is *Waste Management* introduced in Sect. II and shown in Fig. 1a.

Tools & Videos. We created three different versions of GrowingLeaf and multiple training videos (with accompanying handouts) described in Tbl. I and II. To focus the subjects, functionality not needed by the study was removed from the UI in all tool versions. For example, Tool-EI was created so that subjects could use EI-Sim in isolation without exposure to SEI-Sim (row 1 of Tbl. I) and was introduced to the subjects using Video IIEI (row 5 of Tbl. II).

C. Procedure: Conducting the Experiment

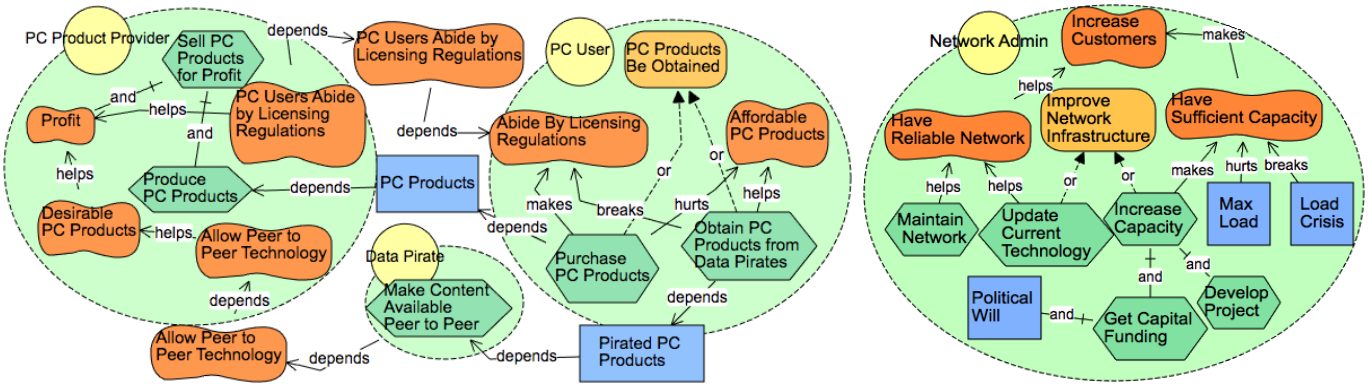
For our study, subjects were required to be graduate students with a basic understanding of requirements engineering and proficiency in English. Subjects were recruited through group mailing lists and an introductory graduate-level course in requirements engineering, and were offered a chance to win a $50 gift certificate. Sixteen subjects volunteered for the study. One subject was unable to complete the study due to time constraints and their data has been excluded from this analysis. 9 Masters and 6 PhD students participated.

The experiment was carried out in November 2016, one-on-one with the subjects in a meeting room. At the start, the subjects were asked to rate their familiarity with requirements engineering, and the iStar modeling language. Fig. 3 shows the subjects' responses: the majority were somewhat familiar with requirements engineering and not at all familiar with iStar. We did not have any subjects who believed they were extremely familiar (or experts) in the field.

[3] http://www.cs.toronto.edu/~amgrubb/archive/RE17-Supplement

[4] Called "Leaf Simulate" in previous versions of the tool.

[5] Called "Stochastic Simulate" in previous versions of the tool.



(a) Trusted Computing model, used in RQ0.

(b) Network Administrator model, used in RQ1.

Fig. 2: Trusted Computing model and Network Administrator model.

TABLE I: GrowingLeaf tool versions created for the experiment.

Name	Rationale	Functionality
GrowingLeaf-EI-Sim (Tool-EI)	For focused learning of EI-Sim and EIs.	Analysis view shows Forward Analysis and EI-Sim. No changes to the Modeling view.
GrowingLeaf-SEI-Sim (Tool-SEI)	Control for SEI-Sim to prevent learning effect of EIs.	Analysis view shows Forward Analysis and SEI-Sim. Function Type selection removed from Modeling view right panel.
GrowingLeaf-Forward Analysis (Tool-FA)	Intro version without EIs or SEIs to prevent a learning effect.	Analysis view shows only Forward Analysis. Function Type selection removed from Modeling view right panel.

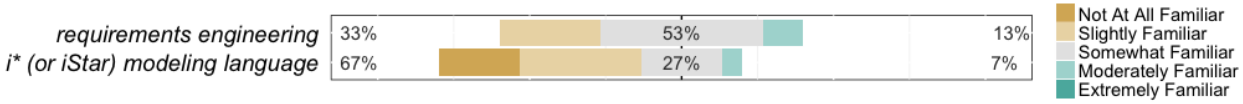


Fig. 3: The subjects' self-reported level of familiarity with requirements engineering and the iStar modeling language. The shown percentages (left-to-right) refer to Not At All Familiar + Slightly Familiar, Somewhat Familiar, and Moderately Familiar + Extremely Familiar, respectively.

TABLE II: Videos created for the experiment.

	Name	Description
RQ0	Video 0A	Reviewed goal modeling concepts/notations & introduced Tool-FA.
	Video 0B	Introduced forward analysis with Tool-FA.
RQ1	Video IEI	Introduced EIs.
	Video ISEI	Introduced SEIs.
RQ2	Video IIEI	Introduced EI-Sim with Tool-EI.
	Video IISEI	Introduced SEI-Sim with Tool-SEI.
	Video IIAFA	Introduced Rep-FA with Tool-FA.

The subjects were randomly placed into one of four *subject groups*: Group A with five subjects (called A1-A5), Group B with five subjects (called B1-B5), and Group C, which was further divided into Group CA with three subjects (called CA1-CA3) and Group CB with two (called CB1 and CB2). The study procedure is listed chronologically in Tbl. III. For example, RQ1 is discussed in the second line where Group A learned EIs by watching Video IEI and using Tool-EI, while Group B learned SEIs by watching Video ISEI and using Tool-SEI. Group CA and Group CB skipped this step to prevent a learning effect. Their completion of RQ1 is listed in the fourth line of Tbl. III. We recorded answers electronically and documented questions asked by the subjects as well as novel uses of the tool/analysis.

TABLE III: Study procedure for each research question. Each triple consists of what the subjects learned/did, which video they watched, and which tool they used.

	Subject Groups			
			Group C	
	Group A ($n = 5$)	Group B ($n = 5$)	Group CA ($n = 3$)	Group CB ($n = 2$)
RQ0	iStar & GrowingLeaf, Video 0A, Tool-FA Forward Analysis, Video 0B, Tool-FA			
RQ1	EIs, Video IEI, Tool-EI	SEIs, Video ISEI, Tool-SEI	*skip*	
RQ2	EI-Sim, Video IIEI, Tool-EI	SEI-Sim, Video IISEI, Tool-SEI	Rep-FA, Video IIAFA, Tool-FA	
RQ1	*skip*		EIs, Video IEI, Tool-EI	SEIs, Video ISEI, Tool-SEI
RQ3	Tool Evaluation, N/A, N/A			

IV. Results

In this section, we describe results for RQ0 - RQ3.

A. RQ0

Remembering and Understanding iStar. To evaluate basic cognition, the subjects answered six questions, and their answers were scored out of 11 total correct answers (some questions asked them to identify more than one intention, e.g., "Name all the actors in the model?"). Fig. 4 contains the bar chart for the subject scores, sorted by subject group. Nine subjects received a perfect score; three had one error; and one subject each received the scores 7–9 out of 11. Our null hypothesis was that the subject groups performed equally well on the questions. Using the *Kruskal-Wallis Rank Sum* (KWRS) test [12] we failed to reject this null hypothesis ($p = 0.96$) meaning that we could not detect a difference between the groups. These results show that *the subjects were able to successfully answer remembering and understanding questions, and are comparable*.

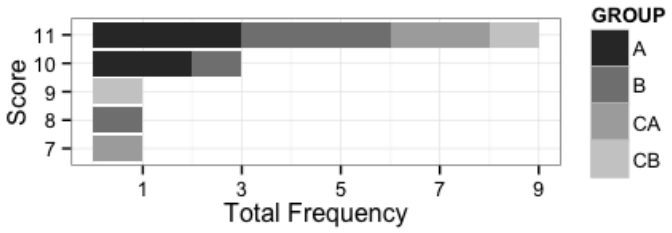


Fig. 4: Stacked bar chart of the subjects' scores in RQ0.

Applying Forward Analysis. The subjects were asked "If the PC Product Provider can only do one of Produce PC Products or Allow Peer to Peer Technology, which one is best for the top goals (use forward analysis to evaluate the alternatives)? Why?". All subjects successfully applied forward analysis. Four subjects chose to satisfy Produce PC Products while eleven chose to satisfy Allow Peer-to-Peer Technology. Subjects who chose Allow Peer-to-Peer Technology gave the justification that it resulted in more root intentions becoming satisfied (more checkmarks). Of those who chose Produce PC Products, two required Abide By Licensing Regulations be satisfied; one focused only on PC Product Provider; and the other made an error remembering previous forward analysis results. Eleven subjects only looked at the analysis output when using forward analysis. *The subjects were successful in the application of forward analysis*.

RQ0 Completion Time. We evaluated the times the subjects took to answer questions, with the null hypothesis being that the subject groups answered questions in a similar length of time. Using the KWRS test, we again failed to reject this null hypothesis ($p = 0.71$), meaning that *the subject groups were not significantly distinguishable in completion times for RQ0*.

> We conclude that the subjects performed similarly on remembering, understanding, and application tests, enabling us to compare the groups.

B. RQ1

Understanding EI and SEI Functions. We asked "If Political Will has *Stochastic* (*R*) as its dynamic type and PD as its current evaluation, what possible evaluations will it have in the future?". All Group A subjects and all but one Group B subject answered correctly. Group A was asked "If Develop Project has *Monotonic Positive* (*MP*) as its dynamic type and PD as its current evaluation, what possible evaluations will it have in the future?". Understanding monotonic positive functions, four answered "PD, PS, FS" (i.e., the correct answer), and four answered "PS, FS", assuming that the value must change between time points. This shows that *the subjects understood the concept of EIs and SEIs*.

Evaluating EIs and SEIs. The subjects in both groups were asked to identify which intentions change over time, and specify how they change. Tbl. IVa lists the recorded answers for each subject, with each row identifying each intention in the model in Fig. 2b, and each column listing whether each subject identified the intention as changing (blank if they didn't) and their specification of that change. For example, Maintain Network was assigned a *Constant* (*C*) function by A1. On average, Group A subjects identified two additional intentions as changing. Both groups identified primarily leaf intentions (as expected), but most subjects believed that intermediate or root intentions changed as well. Group A subjects were able to identify EI functions by name. In describing how intentions change, Group B subjects identified many of the functions defined in EIs. For example, B1 identified Load Crisis as "a sudden change from satisfied to denied", which describes the *Satisfied-Denied* (*SD*) function. B2 identified a new dynamic function where the value is stochastic (similar to *R*) but only between the values *FS* and *FD*. Subjects A3, A4, A5, and B3 wanted to assign a periodic function oscillating between two values. We conclude that *both groups were able to evaluate changing intentions and there was little difference in priming with EIs or SEIs*, when subjects were asked to refine SEIs. We also concluded that *EI functions were intuitive*, because Group B subjects identified EI functions in the absence of priming.

Creating a Representation for Changing Intentions. We asked Group B "for the purpose of communicating with stakeholders how would you represent these dynamics symbolically in the model?". Five subjects recommended using some form of a sparkline. Of those, two recommended using a single sparkline, two recommended showing probability distribution functions with the sparkline, and one recommended showing the intention information in a separate table. CB2 recommended adding a delta symbol to intentions that change, and CB1 recommended blurring the intention borders to differentiate between them. We found *the majority of subjects created sparklines to identify changing intentions*. GrowingLeaf's design already uses sparklines to illustrate EI functions.

Evaluating a Model using Forward Analysis with EI and SEI Priming. We asked subjects "if you can only do one of Update Current Technology or Increase Capacity[6], which one is best for the top goals (use forward analysis to evaluate

[6]While asking this question, the interviewee manually removed intentions to make Increase Capacity a leaf intention.

TABLE IV: Subject data for RQ1 and RQ2.

(a) The subjects' identifications of changing intentions in the Network Administrator model (see Fig. 2b) for RQ1. Entries for each subject list 'F', 'P', 'A', '?', or an EI function (see Sect. II). 'F' defines a function that stochastically changes between *FS* and *FD*. 'P' specifies a periodic function that oscillates between two values. 'A' identifies when the subject determined that an intention changes only as the result of analysis. '?' indicates that the subject either forgot to identify a function or wasn't sure. Position lists leaf & root intentions.

Elements	Position	A1	A2	A3	A4	A5	CA1	CA2	CA3	B1	B2	B3	B4	B5	CB1	CB2
Max Load	leaf	?	*I*	P	*R*	P	*C*			*DS*	F	*R*	*R*	*R*	*I*	*R*
Load Crisis	leaf	?	*R*	*R*	*R*	P	*R*			*SD*	F	*R*	*R*	*R*	*R*	*R*
Political Will	leaf		*R*	*C*	*R*	*R*	*R*			*R*	*R*	*R*	*R*	*R*	*R*	
Update Current Technology	leaf	*MP*	*MP*	P	*R*	*I*			*I*		F	P	*DS*	*R*		*SD*
Maintain Network	leaf	*C*	*MP*	*C*	*R*	*C*					F	*C*	*C*	*R*		
Get Capital Funding	intermediate	?	*MP*	*R*	*I*	P		*R*	*SD*					*R*		
Develop Project	leaf		*MP*	*MP*	*MP*	P			*MP*			*C*	*C*	*I*		
Increase Capacity	intermediate			*R*	*MP*	*I*		D					*UD*	*R*		*DS*
Have Reliable Network	intermediate	*C*		*R*	*R*	A	?							*R*		
Improve Network Infrastructure	root			*R*	?	A		*I*						*R*		
Have Sufficient Capacity	intermediate			*R*	*R*	A		*MP*						*R*		
Increase Customers	root			*R*	*R*	A		*R*						*R*		

(b) Tradeoffs identified by the subjects in the Waste Management model (see Fig. 1a) for RQ2. '*' indicates that the subject considered this tradeoff through Space in Dump. '**' indicates that Produce Green Waste was considered as a tradeoff with one of Use New Dump, Build Large Dump, Manage City Waste, Willingness to Separate Waste, or itself.

Elements	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	CA1	CA2	CA3	CB1	CB2
Build Small Dump / Build Large Dump	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Use Current Dump / Use New Dump	X	X	X	*	X	X	X	*	X	X	X		X	X	
Build Green Centre / Upgrade Trucks			X					X	X			X			X
Produce Green Waste / **				X			X			X				X	X
All (leaf) intentions			X	X										X	
Space in Dump														X	X
Comply with Standards							X								
GW Education Program / Use New Dump								X							
Reduce Operating Costs / Positive City Image													X		

(c) The subjects' responses to the impact of a change in the evaluation of Environmental Concern (see Fig. 1a).

Response Categories	A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	CA1	CA2	CA3	CB1	CB2
No Impact (Not much)												X			
One or Two Steps Graph Trace	X					X			X						X
Static Evaluation of Values			X								X		X		
Multi-Step Graph Trace										X					
Explanation of an Interaction					X		X							X	
Scenario Generation		X		X				X							

the alternatives)? Why?". All subjects answered correctly (i.e., Increase Capacity). Three subjects stated the dominance of the *Makes* relationship between Have Sufficient Capacity and Increase Customers. Twelve subjects used trial and error by assigning various values (of which two did not assign values to other leaf intentions). Subjects A4 and B2 discussed changing intentions in the model, and A5 used EI functions. We conclude that although some subjects considered changing intentions, *priming with EIs and SEIs did not impact the subjects' model evaluations with forward analysis*.

RQ1 Completion Times. Finally, we wanted to see if priming with EIs or SEIs affected the subjects' RQ1 completion times, as well as if there was a fatigue effect for Group C subjects who completed RQ1 last. We tested the time between Group A and Group B, between Group A and Group CA, and between Group B and Group CB. For each of these three tests our null hypothesis was that there was no difference between the groups. Using the KWRS test, we failed to reject this null hypotheses: ($p = 0.25$) for Group A vs. Group B; ($p = 0.88$) for Group A vs. Group CA; ($p = 0.053$) for Group B vs. Group CB. Upon further investigation of our test of Group B vs. Group CB (which was barely significant), we found that both subjects in Group CB completed RQ1 faster than Group B. We conclude *there was no difference in times between priming with EIs or SEIs*. If a fatigue effect existed in Group CB, it may correlate with the reduced number of changing intentions identified by this group.

Model Changes. To elicit possible errors or threats to the analysis, we asked the subjects how they would change the Network Administrator model. Most recommended adding additional relationships to the model. Two recommended expanding Have Reliable Network, and another recommended removing Maintain Network. One subject thought that Improve Network Infrastructure should be a soft-goal while another thought that since it was a hard goal, it should be physically located above Increase Customers in the model. *These recommendations do not impact the results for RQ1*.

The subjects in both groups understood and evaluated EI and SEI functions, and evaluated intentions with them. SEI subjects created functions and graphical representations similar to EIs. EIs were found to be intuitive. Priming with EIs and SEIs did not impact forward analysis.

C. RQ2

Understanding and Analyzing Tradeoffs with Simulation. We asked the subjects to "Identify all the alternative decisions

in the model", and then asked them to use their assigned analysis techniques to "choose the best alternative for each decision". After, we asked them to "describe how the element evaluations vary over your analysis? Do any trends emerge?". Tbl. IVb lists the tradeoffs identified by the subjects. All subjects identified a tradeoff between Build Small Dump and Build Large Dump (first row), and most identified the tradeoff between Use Current Dump and Use New Dump. Five subjects incorrectly identified Build Green Centre and Upgrade Trucks as alternatives. Other alternatives discussed included Produce Green Waste compared with other intentions or itself, by five subjects. Individual items like Space In Dump and Comply with Standards were seen as tradeoffs themselves. B3 compared GW Education Program and Use New Dump, while C3 compared two soft-goals Reduce Operating Cost and Positive City Image.

Three Group A subjects used EI-Sim by selecting dynamics for the model and obtained meaningful results. For example, A1 said "... [set] Space in Dump to *Decrease* (*D*) and now we have a problem. At some point, we will have an issue with the waste. When we have denied, we need to make sure the small or large dump is built". Two subjects used only the *Constant* (*C*) function, reducing the power of their simulations to forward analysis. Four out of five Group B subjects gained a better understanding of the model structure (via relationships) and considered the model as a whole. For example, B3 said "... Process Green Waste really needs a *PS*, because *FS* will not [satisfy] Positive City Image or Reduce Operating Cost". The remaining subject focused on Manage City Waste to make decisions. Group C subjects used a divide and conquer approach to decision making, making one decision at a time. For leaf intentions not included in their current decision, they either applied *FS* or did not apply an evaluation label. For example, CA1 said "satisfying Use Current Dump would only satisfy Manage City Waste, but would not affect the decision for Build Small Dump or Build Large Dump". We conclude that *there is some effect between the use of simulation and model analysis with EI-Sim and SEI-Sim affecting the subjects' understanding of the model structure*. We have mixed results for the effectiveness of EI-Sim. *EI-Sim affected the subjects' understanding of time-based events*, but two subjects did not use EIs. Rep-FA resulted in the subjects not considering the full model, making disjoint decisions instead. *Priming with EIs and SEIs did not have an effect on how the subjects identified trade-offs*. We learned that we need to better connect EI-Sim with EIs to improve the subjects' usage of EI-Sim.

Analyzing a Single Changing Intention. We asked "What would be the impact if Environmental Concern changes in the future?", and categorized each subject's answer, listing the categories in the rows of Tbl. IVc (ordered from worst to best). Responses that considered the impact in more depth were considered to be better. For example, answers that considered the propagation of multiple links were better than those that considered only one link, and considering possible evaluations of Environmental Concern and its impact on other interactions — better still. Two Group A and one Group B subjects generated scenarios for Environmental Concern. The best Group C answer was CB1 who explained the interaction as "it negates the efforts of the GW Education Program. But you still get to Manage City Waste". The data in Tbl. IVc signals that Group A and Group B had responses of more depth than Group C. This evidence was not significant but corroborates our previous result that *simulation improved the subjects' understanding of the model structure*.

Evaluating a Task Ordering Tradeoff with Simulation. We asked Group A/Group B [resp. Group C] the question "Assume you can sequentially complete both Build Green Centre and Build Small Dump. Which order is best for the top goals (use simulation [resp. forward analysis] to evaluate the alternatives)? Why?". Three Group A subjects used EI-Sim and selected Build Small Dump, concluding that it was most important to Manage City Waste and that its satisfaction was dependent on Use Current Dump or Build Small Dump, and there were no guarantees that the current dump would last if the City completes Build Green Centre first. Those in Group A who did not use EIs (or just used constant values) selected Build Green Centre (or Process Green Waste). One made assumptions about the relative length of time it would take to build each item, and the other looked at the impacts on only two of the soft-goals. Group B unanimously chose Build Small Dump, but their justifications varied. Three subjects cited Build Small Dump's contribution to Manage City Waste as the key reason. One subject stated a preference to satisfy the hard goals over the soft-goals. Another subject cited that Build Green Centre required Upgrade Trucks in order to satisfy Process Green Waste, whereas Build Small Dump directly impacted Manage City Waste and Reduce Operating Costs. Three Group C subjects chose Build Small Dump, concluding that since Build Green Centre does not affect Manage City Waste, Build Small Dump is preferable. These subjects did not consider the ordering; instead, they selected Build Small Dump assuming the City could only complete one alternative. CA1 selected Build Green Centre via Process Green Waste by evaluating the number of root intentions impacted by each alternative. CA2 concluded that "it doesn't matter". We conclude that *EI-Sim enabled the evaluation of task ordering tradeoffs. EI-Sim and SEI-Sim helped the subjects understand the model structure. Using Rep-FA, the subjects did not consider time-based information*.

RQ2 Completion Times. We tested to see if there was a difference in completion time between Group A, Group B, and Group C, to see if the simulation type affected RQ2 completion times. Our null hypothesis was that there was no difference. Using the KWRS test, we failed to reject this hypothesis ($p = 0.054$), but since it was arguably significant, we ran Dunn's test [13]. Dunn's test performs post-hoc pair-wise comparisons between groups found significant with KWRS. We found that Group A was significantly different from Group B ($Z = 2.3,\ p = 0.0098$) and Group C ($Z = 1.7,\ p = 0.045$), but there was no significant difference between Group B and Group C ($Z = -0.64,\ p = 0.26$). Group A subjects took an

average of almost six minutes longer to complete RQ2 than the other two groups. We conclude that *Group A's completion times were significantly longer than Group B and Group C, which were not significantly distinguishable from each other*.

Model Changes. As with RQ1, we elicited threats to the analysis by asking the subjects how they would change the Waste Management model. Multiple subjects recommended removing or connecting intentions that had no links (such as Purchase Land), and adding additional links. One also noted that Purchase Land was incorrectly modeled as a resource instead of a task. Other recommendations included quantifying the size of the dumps, making Environmental Concern more explicit, and changing the layout of the model. One subject found the dependum confusing. Finally, one subject in Group B thought that the model needed a long-term city waste management goal. *These recommendations do not impact the results of RQ2*.

EI-Sim and SEI-Sim improved the subjects' understanding of the model structure, and EI-Sim improved the subjects' ability to reason about goal models over time, but this analysis took significantly longer. We learned that SEI-Sim, which was created as a control for comparison with EI-Sim, improved the subjects' understanding of the model. The user-generated simulation (Rep-FA) proved difficult for the subjects to answer time-focused questions.

D. RQ3

Tool Improvements. We asked the subjects "What suggestions or changes would you recommend to the developers of this goal modeling tool?" and list their answers, grouped into required, desired, and other, in Tbl. V[7]. Since the completion of the study, we have implemented all of the Required Improvements and are working on some of the Desired Improvements, all of which need some additional computational analysis to complete. Items listed as Other Recommendations are either outside the scope of our current research plan, have already been implemented but were not visible to the subjects, or are recommendations for the underlying language.

Tool Rating. We asked the subjects to rate the tool based on their level of satisfaction with ease of use, appearance, modeling functionality, and analysis functionality. Fig. 5a contains the likert graphs of the subjects' evaluations. *Ease of use* was evaluated the lowest, with 87% of the subjects satisfied and 13% unsatisfied. The subjects rated the *appearance* of the tool best, with 93% being satisfied and 7% being indifferent. The subjects rated the *analysis functionality* better than the *modeling functionality*. We also asked the subjects "How likely is it that you would recommend this goal modeling tool to a colleague?" The likert graph of the subjects' responses is shown in Fig. 5b. All but one subject thought it was likely that they would recommend GrowingLeaf. We conclude that *the subjects were overwhelmingly satisfied with the tool*.

[7]When the same suggestion was made by multiple subjects, we indicate it by 'x2' for two subjects and 'x4' for four subjects.

TABLE V: GrowingLeaf Improvements.

Required Improvements
- Clear all intention evaluation labels. (x4) - Clear all dynamic function labels. - Disable delete key. - Allow intention names to span multiple lines. - Indicate whether something has changed in the previous step. - Add legend for the dynamic function labels and evaluation labels. - Change length of slider depending on the type of analysis. Make slider for forward analysis shorter (i.e., two steps wide). - Make PS/PD dots more obvious. - Create onscreen Help with instructions.
Desired Improvements
- Syntax checking while the user is modeling. - Highlight and unhighlight leaf intentions. (x2) - Highlight and unhighlight root intentions. - Overlay and or compare simulation paths. - Auto-resize model and intentions based on font size. - See steps within a forward analysis propagation.
Other Recommendations
- Change Depends arrows to make the arrow go in the opposite direction. (x2) - Make the shapes for goals and soft-goals more distinctive. - Prevent users from assigning dynamics to non-leaf intentions. (x2) - Create analysis where some values are fixed and others vary. - Add a cost or utility function for each decision and then automate the analysis to figure out the optimal solution.

Tool Version Comparison. We examined if there was any difference in ratings across tool versions. Our null hypothesis was that there was no significant difference in how the subject groups rated the tool (ease of use, appearance, modeling functionality, analysis functionality, and likelihood to recommend the tool). Using the KWRS test, we failed to reject this hypothesis ($p = 0.81, 0.83, 0.80, 0.24, 0.81$ respectively), meaning that *there was no discernible difference between the evaluations of the subject groups*.

The subjects rated GrowingLeaf highly and found it usable.

V. Discussion

This section discusses our statistical methods, followed by the broader implications and threats to validity of our study.

A. Statistical Methods

We use nonparametric statistics (specifically, the KWRS test) to evaluate if there are distinct groupings within our sample data. Nonparametric statistics holds two main advantages over its parametric counterpart. First, for small sample size numerical data, nonparametric statistics avoids being unduly influenced by data points which differ greatly in magnitude. The completion time data benefited from this reduced sensitivity. Second, nonparametric statistics imposes no assumptions about the underlying shape of the probability distribution in its computation, and all our data benefited from this. Yet, our small sample size still limits the power of these tests.

B. Implications for Research

The subjects were able to use EIs and EI-Sim, validating our approach for goal modeling over time, but their use was not perfect. The subjects missed the nuanced differences between the *Monotonic Positive* (*MP*) and *Increase* (*I*) functions, and thought the value should change at each time point. We need to

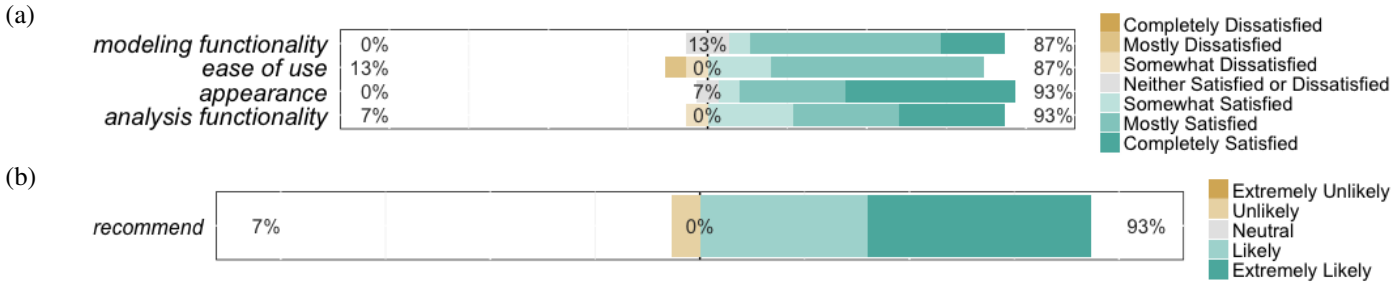


Fig. 5: The subjects' evaluation of GrowingLeaf. (a) Likert graphs of the subjects' level of satisfaction with modeling functionality, ease of use, appearance, and analysis functionality. The shown percentages (left-to-right) refer to Completely + Mostly + Somewhat Dissatisfied, Neither Satisfied or Dissatisfied, Somewhat + Mostly + Completely Satisfied, respectively. (b) Likert graph of the subjects' likelihood to recommend GrowingLeaf. The shown percentages (left-to-right) refer to Extremely Unlikely + Unlikely, Neutral, Likely + Extremely Likely, respectively.

understand further why not all Group A subjects used EI-Sim effectively, and better describe how EI-Sim depends on EIs. In some parts of the study the subjects paid closer attention to the content of the model than in others. We believe this to be a result of the lab setup and asking questions in isolation, but note that asking the "Why?" questions gave us access to how the subjects were thinking.

C. Implications for Education

Our experiment was done with iStar learners so it has implications for teaching. We used the dimensions of Bloom's taxonomy to assess subjects cognition at multiple levels. We observed that subjects had difficulty with the *Depends* link and why it propagates information in the opposite direction of the link; the fact that the *Breaks* contribution link propagates a *PS* from *FD*; when to use the *Unknown* (*U*) label; and which intentions to focus on in trade-off analysis. This is consistent with previous reports discussing students' issues with *Depends* links and analysis [14]. Our study found that SEI-Sim helps subjects understand the structure of the model and how links propagate. Since SEI-Sim can be used without EIs, we believe it can be helpful in teaching propagation rules because learners see many possible analysis combinations and are able to ask questions about propagation results.

D. Threats to Validity

We discuss threats to validity using the categories in [15].

Conclusion Validity. Our main threat is our low sample size. With only fifteen subjects, the statistical power of any relevant statistical test will be low but we believe that the tests used were appropriate for our data (see Sect. V-A). We automatically recorded completion times, starting immediately after each post-video discussion, to ensure reliable measurements. Data collection and analysis was made independent to reduce researcher bias. To mitigate *reliability of treatment implementation*, we standardized the experiment by maintaining the experiment setup throughout the study period, and used videos and handouts to ensure that the subjects had the equivalent training material (see Sect. III). We do not believe there is a *random heterogeneity of subjects* risk in our study since our population was homogeneous, having similar knowledge, abilities, and previous experience with iStar and RE. Collecting addition demographics information would provide further evidence. Although researchers personally knew some of the subjects, they did not discuss the study with them prior to conducting it.

Internal Validity. We used timing data and understanding questions to check for and mitigate against *selection-maturation interactions*, where one group learns a treatment faster than another. We gave subjects time to review training materials to ensure they were ready to answer questions. We also checked for a *maturation effect* in RQ1, which Group C completed after RQ2, and we found an effect in the results of Group CB (discussed in Sect. IV). Since study participation was voluntary, a *selection effect* may exist because only motivated subjects participate. Further replication with an entire graduate class could mitigate this effect. To our knowledge, no subjects had used GrowingLeaf prior to the experiment, or iStar outside of course requirements.

Construct Validity. This study was specifically designed to validate EIs as a construct and we feel they were accurately represented. In evaluating EI-Sim, we included both SEI-Sim and Rep-FA to evaluate whether there was a *mono-operation bias* because the SEI construct, by definition, is an under-represented version of EIs. We asked questions across dimensions of Bloom's taxonomy to mitigate against *mono-method bias*. We mitigated *experimenter expectancies* by asking questions explicitly as worded in our protocol. As always, we have threats of *hypothesis guessing* and *evaluation apprehension*. Multiple subjects noted being nervous about the study because they were still quite novice. Additional studies evaluating these constructs can mitigate these threats.

External Validity. Our homogeneous population (see Conclusion Validity) means that we cannot generalize our findings to the broader population of modelers. Further experiments with different populations, problem domains, and larger models for scalability are required to generalize these results. Our study was conducted one-on-one in a lab, which provides a foundation for further case studies, but does not directly generalize to early-phase requirements engineering done in

"real" groups. We simulated the evolution of models in this study. Since our design was not longitudinal in nature, the subjects did not have the ability to witness changes in intention evaluations.

VI. Related Work

Here, we compare our experiment with related work.

Modeling Tools and Techniques. Extensions and tooling for iStar models have been previously developed [16]. OpenOME [17] was developed to evaluate forward and backward analysis algorithms and was frequently used in iStar case studies, [18]–[20]. Pimentel et al. presented a tool focused on the creation of State Charts from goal models [21]. Amyot et al. combined Use Case Maps with Goal-oriented Requirement Language (GRL), a standardized version of iStar [22]. GRL has been extended for legal compliance [23] and most recently to add changing intentions [24]. The approach presented in [24] differs from ours in that it discusses changing intentions with quantitative evaluations in absolute time using OCL constraints. It has not been validated beyond the illustrative example. Creative Leaf [25] added creativity techniques to goal modeling, and shares foundational code with GrowingLeaf. Further studies of Creative Leaf will complement the usability aspect of our work, but since Creative Leaf does not contain changing intentions, its studies will not interact with ours.

Methods. We built on the methodology of similar studies in RE for our between-subjects experiment and followed the guidance in [15], [26]. Karras et al. reported on a between-subjects lab experiment similar to ours [27]. Their treatment group had significantly longer completion times than their control group, and we suspected that this was due, in part, to only one group learning a new tool. We attempted to control for this by giving the subjects exploration time with GrowingLeaf prior to asking them questions. Santos et al. reported a within-subjects quasi-experiment evaluating layouts of iStar models [20]. A within-subject experiment was not possible for our study due to a learning effect. We removed the defect detection task from our study after considering the drawbacks discussed by Santos et al. and the similarities in our subject populations.

VII. Conclusions and Future Work

In this paper, we presented a between-subjects experiment to evaluate EIs, EI-Sim, and GrowingLeaf. We concluded that the representation of EIs was suitable to the task of identifying and representing intentions over time, that EI-Sim improved the subjects' ability to reason about goal models over time, and that GrowingLeaf was found to be effective and usable.

The study results are encouraging, and we received useful feedback. Future work will focus on improving the study, generalizing our findings to a broader population (including industry-based modelers), and investigating how to resolve modelers' disagreements over EI function choices for intentions. Future work will also replicate these results with other subjects and domains, and evaluate other analysis techniques within our tool. As our tool gains maturity, we would like to evaluate EIs with a longitudinal study where stakeholders might actually observe their models evolving.

Acknowledgments. We thank Fabiano Dalpiaz, Jeffrey S. Castrucci, and members of the Modeling group in Toronto for helping improve this work.

References

[1] E. Yu, "Towards Modeling and Reasoning Support for Early-Phase Requirements Engineering," in *Proc. of RE'97*, 1997, pp. 226–235.
[2] D. Amyot, "Introduction to the User Requirements Notation: Learning by Example," *Comput. Netw.*, vol. 42, no. 3, pp. 285–301, Jun. 2003.
[3] P. Giorgini, J. Mylopoulos, and R. Sebastiani, "Goal-oriented Requirements Analysis and Reasoning in the Tropos Methodology," *Eng. Appl. Artif. Intell.*, vol. 18, no. 2, pp. 159–171, 2005.
[4] A. van Lamsweerde, *Requirements Engineering - From System Goals to UML Models to Software Specifications*. Wiley, 2009.
[5] J. Horkoff and E. Yu, "Interactive Goal Model Analysis For Early Requirements Engineering," *Requir. Eng.*, vol. 21, no. 1, 2016.
[6] A. M. Grubb, "Adding Temporal Intention Dynamics to Goal Modeling: A Position Paper," in *Proc. of MiSE@ICSE'15*, 2015.
[7] A. M. Grubb and M. Chechik, "Looking into the Crystal Ball: Requirements Evolution over Time," in *Proc. of RE'16*, 2016.
[8] A. M. Grubb, G. Song, and M. Chechik, "GrowingLeaf: Supporting Requirements Evolution over Time," in *Proc. of i* Wrksp'16*, 2016.
[9] J. Horkoff and E. Yu, "Comparison and Evaluation of Goal-Oriented Satisfaction Analysis Techniques," *Requir. Eng.*, vol. 18, no. 3, 2013.
[10] D. R. Krathwohl, "A Revision of Bloom's Taxonomy: An Overview," *J. Theory Into Practice*, vol. 41, no. 4, pp. 212–218, 2002.
[11] J. Horkoff and E. Yu, "A Qualitative, Interactive Evaluation Procedure for Goal-and Agent-oriented Models," in *Proc. of CAiSE'09*, 2009.
[12] R. P. Runyon, *Nonparametric Statistics: A Contemporary Approach*. Addison-Wesley, 1977.
[13] O. J. Dunn, "Multiple Comparisons Using Rank Sums," *Technometrics*, vol. 6, no. 3, pp. 241–252, 1964.
[14] J. Horkoff, J. Lockerbie, X. Franch, E. S. K. Yu, and J. Mylopoulos, "Report on iStarT'15," *ACM SIGSOFT SEN*, vol. 40, no. 6, 2015.
[15] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*. Springer Berlin Heidelberg, 2012.
[16] "i* wiki: i* tools," http://istar.rwth-aachen.de/tiki-index.php?page=i%2A+Tools, 2015, accessed: 2015-02-20.
[17] J. Horkoff, Y. Yu, and E. Yu, "OpenOME: An Open-source Goal and Agent-Oriented Model Drawing and Analysis Tool," in *Proc. of i* Wrksp'11*, 2011, pp. 154–156.
[18] J. Horkoff, "Iterative, Interactive Analysis of Agent-Goal Models for Early Requirements Engineering," Ph.D. dissertation, University of Toronto, 2012.
[19] N. Niu, A. Koshoffer, L. Newman, C. Khatwani, C. Samarasinghe, and J. Savolainen, "Advancing Repeated Research in Requirements Engineering: A Theoretical Replication of Viewpoint Merging," in *Proc. of RE'16*, 2016, pp. 186–195.
[20] M. Santos, C. Gralha, M. Goulão, J. Araújo, A. Moreira, and J. Cambeiro, "What is the Impact of Bad Layout in the Understandability of Social Goal Models?" in *Proc. of RE'16*, 2016, pp. 206–215.
[21] J. Pimentel, J. Vilela, and J. Castro, "Web tool for Goal Modelling and Statechart Derivation," in *Proc. of RE'15*, 2015, pp. 292–293.
[22] D. Amyot, A. Shamsaei, J. Kealey, E. Tremblay, A. Miga, G. Mussbacher, M. Alhaj, R. Tawhid, E. Braun, and N. Cartwright, "Towards Advanced Goal Model Analysis with jUCMNav," in *Proc. ER'12 Wrkshps*, 2012, pp. 201–210.
[23] S. Ghanavati, A. Rifaut, E. Dubois, and D. Amyot, "Goal-Oriented Compliance with Multiple Regulations," in *Proc. of RE'14*, 2014.
[24] Aprajita and G. Mussbacher, "TimedGRL: Specifying Goal Models Over Time," in *Proc. of MoDRE'16*, 2016.
[25] J. Horkoff and N. A. M. Maiden, "Creative Leaf: A Creative iStar Modeling Tool," in *Proc. of i* Wrksp'16*, 2016, pp. 25–30.
[26] F. Shull, J. Singer, and D. I. Sjøberg, *Guide to Advanced Empirical Software Engineering*. Springer-Verlag New York, Inc., 2007.
[27] O. Karras, S. Kiesling, and K. Schneider, "Supporting Requirements Elicitation by Tool-Supported Video Analysis," in *Proc. of RE'16*, September 2016, pp. 146–155.