Reflection on Evolutionary Decision Making with Goal Modeling via Empirical Studies

Alicia M. Grubb
Department of Computer Science
University of Toronto, Toronto, Canada
amgrubb@cs.toronto.edu

***Abstract*—Goal models have long been used in academia without wide spread adoption in industry. If the fundamental purpose of goal models is to allow stakeholders to generate scenarios and ask "what if" questions, then which parts of the process of model construction, analysis, and evolution benefit from and which are hindered by manual activities? The recent expansion of goal modelling to ask time-based questions further amplifies this issue because significant additional information is required from stakeholders. Through a series of empirical studies, we aim to isolate the processes of model construction, analysis, and evolution for the purpose of studying the utility of goal-oriented requirements engineering approaches and exploring which tasks are essential practices that stakeholders must complete themselves to gain modeling benefit, and which tasks can be simplified through automation. In this process, we will also measure the benefits of completing relevant goal modelling activities with and without timing analysis. In this short communication, we describe our objectives for understanding the benefits of and barriers to goal-oriented requirements engineering.**

I. Introduction & Motivation

Goal-oriented requirements engineering (GORE) helps stakeholders elicit and analyze the goals of projects and intentions of stakeholders. Within the GORE community, many approaches and notations have been developed, for example, NFR [1], i* [2], GRL [3], Tropos [4], and KAOS [5]. The principal artifact of GORE approaches is the *goal model*, which is a visual representation of functional and non-functional requirements of the project with or without an explicit notion of a stakeholder. While some approaches have a textual representation (e.g., iStarML [6], GBRAM [7]), the process of goal modeling is dominated by the creation and analysis of the visual representation. In the early-phases of a project stakeholders are focused on their intentions for the project (i.e., their goals), the system-level requirements, and how the system will integrate into the domain and depend on its environment.

Once a goal model is constructed, stakeholders can generate scenarios and ask "what if" questions. The analysis procedures of each approach mentioned above vary, however there are several trends that have emerged, such as treating the model as a directed acyclic graph for the purpose of analysis. To ask a "what if" question, modelers make evaluation assignments to leaf-level nodes and then propagate evaluations to root-level nodes. This approach is generally called *forward propagation* or *forward analysis*. To generate a possible scenario that leads to the assignment of a set of desired goals (i.e., "is this possible"), stakeholders make evaluation assignments to root-level nodes and then propagate evaluations to leaf-level nodes. This approach is generally called *backward propagation* or *backwards analysis*. Horkoff and Yu provide a detailed comparison of these approaches [8]. They primarily vary in whether the evaluation assignments are qualitative or quantitative data (or both), and the level of automation in the propagation procedure. As in [8], we omit KAOS from this comparison because of the dissimilarities in analysis procedures; however, we will address the applicability of this work for evaluating KAOS in Sec. III-F.

More recent approaches have looked at goal evolution and simulating evaluations over time. Dalpiaz et al. looked at how goals evolve as they are connected with runtime entities [9]. Our prior work provided qualitative analysis through simulation allowing stakeholders to consider alternatives over time and answer time-based questions [10]. Our approach, later called Evolving Intentions, uses automated analysis and enables stakeholders to visualize single simulations and predict outcomes [11]. Aprajita et al. introduced TimedURN, an approach that uses quantitative analysis to capture impacts of trade-offs at future time points [12]. TimedURN provides support for creating different scenarios and visualizing trends over time, and is fully integrated into a previously validated tool [13].

Motivating Example: Adding Bike Lanes to a Major Street. Urban roads are often overcrowded leading to collisions between motorists and cyclists. Proponents of cycling argue for the addition of separated cycling lanes (called bike lanes). We consider a project to add bike lanes to a major residential and commercial thoroughfare in a large Canadian city. The city is considering the impacts of these changes on pedestrians, cyclists, motor vehicles, transit, and other road operations. The City planners are focused on two issues: Are bike lanes or parking most effective on curb-side? Is a permanent solution or a temporary solution most appropriate?

With a goal-oriented approach, (A1) we assume that modelers can construct a goal graph (or model) as shown in Fig. 1. In the partial goal model of the project (see Fig. 1), the stakeholders want to satisfy Have Bike Lanes while satisfying Cyclist Safety and Access to Parking (i.e., some of the *root-nodes* in the graph). Next, (A2) goal-oriented approaches assume that modelers can use analysis to answer trade-off

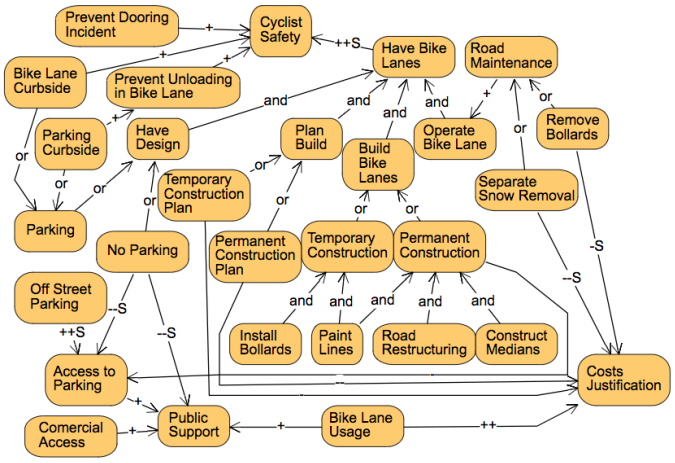

Fig. 1: Goal Model of the Bike Lanes Motivating Example.

questions. An example "what if" question that stakeholders can answer using goal model analysis is, if bike lanes are permanently installed, will Have Bike Lanes, Cyclist Safety, and Access to Parking be satisfied? Finally, (A3) time-based goal modeling approaches assume the evaluations of some elements in a goal model change over time and that goal modeling can be used to express these changes and answer time-based questions. For example, Have Design will become more satisfied at some point, but Bike Lane Usage may vary depending on external factors, such as seasonal changes in weather. Permanent Construction will affect Access to Parking while the roadway is being reconfigured, whereas Temporary Construction will not have the same effect. The city planners want to understand how changes in these goals impact their decisions by exploring two questions: How do variations in the satisfaction of Bike Lane Usage over time affect the City's goals? Given seasonal constrains and roads, what happens to the City's goals if we put bike lanes (or parking) curbside? Both TimedURN and Evolving Intentions argue it is feasible for stakeholders to create visualizations answering these questions.

While GORE methods are successful in a number of ways, these assumptions (A1–3) have not been sufficiently addressed and leave some questions unanswered. For example, what benefit is gained from drawing the initial goal model in Fig. 1? How do stakeholders know their model is valid? How do stakeholders know what analysis method to choose? Do stakeholders use the analysis results to justify their preferred answer or discover the best approach? How can stakeholders verify results of time-based analysis to ultimately find out if results were meaningful? In this work we reflect on some of these assumptions, and ask if they are valid or what impact they have on the utility of the approach overall. We suspect that these assumptions may underlie the limited adoption of GORE approaches.

Barriers to Adoption. Although goal modeling has long been used in the literature, examples of its adoption in industry are limited. Others have surveyed the barriers to adoption of software engineering research more generally. Lo et. al. found evidence that practitioners did not adopt SE research because research tools were not directly applicable to work tasks and the benefits did not outweigh the costs [14]. Ongoing work by Franch et al. is looking at how practitioners perceive RE research [15]. We hope their findings, once published, will document additional opportunities for encouraging GORE adoption. Mavin et al. surveyed experienced requirements engineering practitioners to investigate the lack of adoption of goal models [16]. They found that significant barriers to adoption among practitioners included a lack of tooling, expertise, and organizational enthusiasm, as well as a belief that the approaches do not scale or have sufficient benefits. In their survey of the literature they found that the vast majority of the validation in goal modeling papers focused on feasibility, with very few papers discussing practicality or utility. Furthermore, of the few controlled or quasi-experiments published in the goal modeling community, most have completely drawn models (e.g., [11], [17]) or starter models (e.g., [18]).

One effort already underway to investigate the applicability of i* is Abad et al.'s study of junior consultants' use of SD diagrams in the context of the DHARMA method [19]. They found that practitioners understood the concepts of actors and dependencies, but found issues with ambiguous semantics, and confusing graphical representation. Abad et al. recommended a set of guidelines, such as creating a catalog of actor templates and only considering dependency relationships between two actors at a time, but did not explicitly focus on utility.

Rather than devote significant effort to developing a professional tool, we assert that we must understand where to focus development efforts to mitigate these barriers and encourage adoption. In this work, we consider activities in GORE and question the assumptions that the approach makes, in order to identify the activities that do and do not benefit stakeholders and give recommendations for areas where automation would help increase adoption.

Contributions. In this short communication, **we contribute a protocol to investigate the utility of activities in GORE**. We look at the process of goal modelling through the division of construction, analysis, and evolution. For each activity, we ask two research questions with the goal of understanding where the essential human practises are in GORE: To what extend do stakeholders gain utility in completing the activity? If there is limited utility to be gained, can simplification or automation of the activity provide similar utility with reduced effort?

In the remainder of this paper, we introduce the objectives and assumptions of each activity (see Sec. II), and present a discussion of our study design (see Sec. III). We conclude and call for collaborations in Sec. IV.

II. Activities & Assumptions

In this section, we introduce assumptions made in GORE activities (i.e., construction, analysis, and evolution) and present potential research questions to examine assumptions.

Our primary goal is to measure the 'utility' of each stage of goal modeling. Utility is either viewed as 'fitness for some desirable purpose or valuable end' or an 'intrinsic property

of anything that leads an individual to choose it rather than something else' [20]. In this work, we are interested in the intrinsic utility of GORE rather than utility over other RE approaches (e.g., User Stories). We aim to compare the utility of completing activities in different ways.

We begin by considering where in the GORE approach (i.e., in which activities) is the central utility. Much of the work on analysis and extensions of goal modeling languages (e.g., extensions for security, preferences, legal, regulations, adaptions, and evolution [21]) takes the construction of models as a default, and makes the assumption that the majority of the utility in GORE approaches is in analysis activities. However, in Horkoff and Yu's comparison of forward analysis techniques (i.e., i*, GRL, and Tropos), they found that each technique resulted in a different answer and recommended that the analysis techniques be used as heuristics for model exploration and debugging rather than decision making [22]. Understanding the utility of construction and analysis will be helpful for GORE adoption, and is also required in order to make any claim about the utility of the evolution activities.

A. Construction

We pose the following research questions in order to motivate our work and explore assumptions in construction:

- To what extent do participants gain value in manually drawing goal models (in a tool or on paper)?
- Can participants gain similar benefits from reviewing and correcting a model?
- What benefits do visual representations have over textual representations? To what extent is a visual representation beneficial?
- What other requirements practises are complementary to goal modeling or have similar benefits?
- To what extent is making changes or correcting a model sufficient to understanding them?
- How do these benefits change when there is a single stakeholder vs. groups of stakeholders?

In the motivating example, we can ask: What benefit is gained from drawing the initial goal model in Fig. 1? While this is not an exhaustive list of research questions, we use it to generate two hypotheses: **(HC-1)** When constructing goal models, there is no utility in drawing them on paper. **(HC-2)** When constructing goal models, there is no utility in drawing them in a tool. Through experimentation, we could reject both HC-1 and HC-2 and discover utility in both approaches. If we fail to reject both HC-1 and HC-2 then the central utility in GORE may be in analysis and there may be opportunities for automation of construction activities. Also, if we reject HC-1, fail to reject HC-2, and find utility in analysis then future work may look into using computer vision techniques to translate paper models into computer models for analysis.

To study HC-1 and HC-2, we define three treatments: **(TC-Paper)** subjects construct and review goal models on paper; **(TC-Tool)** subjects construct and review goal models in a tool; and **(TC-Auto)** subjects specify sufficient information to enable for the construction of a goal model and stakeholders review the constructed goal model. These three treatments allow us to explicitly learn what utility there is in each process but by comparing TC-Paper and TC-Tool with TC-Auto, we can understand if simplification or automation can provide similar benefits. While other hypotheses and treatments may exist, we use these in our study discussed in Sec. III-C.

B. Analysis

Measuring the utility of GORE analysis is entirely dependent on the type of analysis chosen for the measurement. In this subsection, we focus on some assumptions that are general and exist across approaches. These assumptions may apply to both the initial analysis in evolutionary approaches as well as the specialized analysis in domain specific extensions (e.g., security, law). We propose the following research questions to examine analysis assumptions and motivate our work:

- What insights do participants gain by assigning initial evaluation labels (or labels for analysis methods)?
- How sufficient is a textual representation for capturing initial evaluation labels (or labels for analysis methods)?
- What utility is there in using analysis for answering questions, debugging, and model comprehension?
- What utility does each technique provide (e.g., forward analysis, backward analysis, metrics over models [23] [24])?
- For each technique, to what extent are there benefits in performing analysis manually vs. automatically?
- To what extent do stakeholders have to understand the underlying analysis to benefit from it?
- How likely is it that stakeholders use the analysis to justify their preferred answer rather than the best result?
- To what degree is there utility in completing analysis manually on paper?

In the motivating example, we can ask: To what extent was asking "what if" questions useful in discovering goals and relationships that were missing from the model? To understand analysis deeper, we propose three hypotheses: **(HA-1)** There is no utility in using analysis techniques for debugging and model comprehension activities. **(HA-2)** There is no utility in using analysis techniques for answering questions. **(HA-3)** There is no additional utility in using interactive analysis techniques over automated techniques. HA-3 specifically focuses on understanding automation opportunities in analysis. When examining these hypotheses, evidence and conclusions gained from studying one technique may not be valid with other techniques. Since there is no unified GORE process and individuals are left to interpret how to proceed in analysis, it is unclear whether stakeholders generate their questions as they are reviewing the model or know them a priori. In the motivating example, did the stakeholders intend to explore the impacts of Permanent Construction, or was it only a result of making connections from the model? The assumption is made that stakeholders can translate their problem into a form appropriate for goal model analysis techniques. How does the City in the Bike Lanes example go from being curious about

Permanent Construction to formulating a forward analysis appropriate input? With this additional assumption, we consider two treatments for evaluating these hypotheses: **(TA-Oz)** subjects ask questions in natural language and are returned results; and **(TA-Self)** subjects are trained on the analysis method and must ask questions and perform analysis independently. The treatments and hypotheses are further described in Sec. III-D.

C. Evolution

Evolution activities assume that not all changes are known in the early stages of a project; thus, stakeholders need to be able to update models and deal with unanticipated changes. We define an evolution activity to be any activity that involves reviewing goal models or analysis results for the purpose of updating/reviewing past decisions or monitoring current conditions. With this definition, we consider the following research questions as motivation:

- To what degree is reviewing past analysis results helpful in understanding current issues?
- To what degree is reviewing models helpful in making decisions in the present?
- To what degree is making projections about the future meaningful for project outcomes?

In the motivating example, we can ask: To what extent is it helpful to make predictions about how variations in Bike Lane Usage will affect other goals? We propose two initial hypotheses to explore evolution activities: **(HE-1)** There is no utility in monitoring goal model elements later in the project lifecycle. **(HE-2)** There is no utility in reviewing past analysis results in understanding current decisions. Addressing HE-1 is outside the scope of our study and should be considered through the lens of action research. If we reject HE-2, then there are automation opportunities for storing and retrieving past models, analysis results, and conclusions. We propose two treatments to evaluate HE-2: **(TE-Model)** subjects are given past goal models and analysis; and **(TE-Info)** subjects are given past questions and decisions. We discuss how these treatments are used in our study design in Sec. III-E.

III. Study Design

In this section, we propose a series of empirical studies aimed to study the validity of assumptions made by the GORE community with respect to the utility of the approach.

A. Overarching Tenets

Prior to introducing our study protocol, we discuss overarching tenets that guide our design.

Stakeholders' Perceptions. Utility, as a construct, is a perceptual judgement of each stakeholder. In order to judge the true utility of something, it must not be evaluated in a hypothetical context. It is not necessarily the case that all studies must be completed *In vivo*. We argue that *In vitro* studies in a laboratory can produce valuable results if stakeholders bring their own problem or project into the lab. Stakeholders must be domain experts in the problem they model and provide their own analysis questions.

Comparing Perceptions. It is difficult to compare individual perceptions of utility. In *between-subject* experiments, we can establish baselines of expertise and experience for comparing subjects and have them perform identical tasks; however, we cannot make claims that subjects would have perceived alternative treatments the same. One mitigating suggestion is to have teams of subjects split into treatment groups and then compare activities of the group as a whole. In *within-subject* experiments, we can give both treatments to the same subject; however, we cannot make guarantees that the subject's perception will not be confounded by the differences between tasks. One mitigating suggestion is to have subjects with large projects split the project into equal sub-projects for each treatment, but this doesn't mitigate the presence of a learning effect. The silver lining in our study is that we are not yet trying to prove that a specific utility exists. We are focusing on discovering potential benefits, so false positives, while undesirable, may be tolerated, and false negatives can best be mitigated through replications.

Remove Modelers. In order for goal modeling to reach wide spread adoption, we need to remove expert modelers from the process. Stakeholders should be able to gain value from the activities without the presence of expert modelers. Studies may need to have a training portion where subjects are given background to engage in goal modeling tasks but should not have expert modelers completing tasks. The exception to this tenet is if expert modelers are used in a 'Wizard of Oz' experiment [25] to mimic some automated process. If utility is found without the immediate presence of expert modelers, then there may be value in automating modeling and analysis processes, as well as investigating how to use the *modeling as a service* paradigm. We would also need to investigate the degree to which stakeholders trust any automation.

Scale Matters. Goal model studies use small models. The utility of the GORE approach could be dependent on model size. Perhaps construction of large models is hard, reducing the utility, but automated analysis of large models would be useful if tooling was sufficiently scaleable. We need to investigate how the value of automated processes depends on model size, as well as the interaction between model size and the cognitive load required to understand models as a whole and the results of analysis procedures. Results from all of these studies must be considered in the context of model size.

B. Common Design Components

In Sec. II, we introduced the dependencies across GORE activities. In this subsection, we give an overview of our study protocol and note common elements between studies. Our protocol consists of ordered study components for each activity: construction, analysis, and evolution. The study for each activity (as described in Sec. III-C through Sec. III-E), taken together, will give us a full picture of the utility of GORE.

For consistency between study replications and to enable longitudinal studies, we plan to collect common data. At the

beginning of each study, we must document, in the subjects' own words, a description of the problem and domain which will be explored with the GORE approach. By documenting this up front, it can be used in later phases of the study as well as in meta analysis. We recommend using a common template to structure the subjects' descriptions. Prior to analysis observations, subjects must be asked whether they have sufficiently modeled their problem/domain? If not, what is missing and why? When subjects are evaluated across multiple activities, subjects would be asked which activity they found as having the most value in understanding their problem/domain. In order to evaluate evolutionary questions, all models and results must be collected and stored. The stakeholders' reasoning and final decision would also be documented.

C. Construction

For studying construction (i.e., HC-1 and HC-2), we propose two complementary approaches. The first study is to evaluate the construction of models within subjects (*Const-Study-Within*). This is most appropriate when a single stakeholder with a large project to model is found. The project is then divided up into three parts, and each part is applied to a treatment (i.e., TC-Paper, TC-Tool, and TC-Auto) in random order. This can be done as a single case study or as a quasi-experiment with a set of stakeholders with independent projects. It is important that the modeler reflects on each treatment separately prior to comparing treatments. The second study is to evaluate the construction of models between subjects (*Const-Study-Between*). This can be completed as a controlled experiment where a baseline can be achieved about the subjects' domain expertise and modeling abilities (for comparing TC-Paper and TC-Tool). The second study is recommended when groups of stakeholders are working on the same project or problem. The group can be randomly divided into one of the three treatment groups to draw separate models. Each subject can reflect on what insights and value the treatment provided and then as a team they can compare and merge their models. While an investigation of the model merge process would be beneficial, it is currently outside the scope of this work.

We chose these designs over focusing on a single treatment and asking subjects what they found useful, as is done in many case studies. This design mitigates *hypothesis guessing* and reduces subjects' desire to please the researcher. This study protocol is process agnostic. We are not advocating the use of any method for automatically constructing goal models, but rather explore the potential benefits. We envision TC-Auto being implemented either by having GORE experts construct the models (i.e., Wizard of Oz), or using natural-language processing or other AI techniques to construct them.

D. Analysis

For initial studies of the utility of analysis, we propose a single within subjects (*Analysis-Study-Initial*) study, where each subject completes both treatments (i.e., TA-Oz and TA-Self). This is not an experiment, but will act as an exploration to the utility of analysis. Each question the subject wants to ask using the model and analysis technique would be documented. Then, depending on the question, an appropriate analysis technique is chosen and the analysis result is produced. The subject is then asked a series of questions to assess the assumptions of analysis: Did the analysis result in changes or improvements to the model? Did the subject successfully answer the question they originally posed, if so, how many analysis iterations did it require?

This study will not give conclusive evidence for HA-1, HA-2, or HA-3, but it will give insights into how to directly test these hypotheses. When subjects complete both treatments (i.e., TA-Oz and TA-Self), deception can be used to make the subjects believe that the goal of the study is to compare techniques, mitigating any performance bias, since the study collects data for both. By asking the subjects to compare techniques, we can investigate what barriers exist but also how deeply the subjects engaged with the analysis and whether they find it believable. We also plan to compare our observations of the results with the subjects' analysis of them. This study can be replicated with individual analysis techniques.

E. Evolution

In an exploratory evaluation of HE-2, subjects may be recalled after several weeks or even months have passed. Subjects are asked to review the question and domain they originally considered and to recall what conclusions they reached. We then reveal the subjects' original conclusions. Subjects are then placed into one of the two treatments groups (i.e., TE-Model or TE-Info) and asked to repeat a modified treatment from either a construction or analysis activity. Where two time-delayed interventions are possible, subjects can be evaluated with both treatments (i.e., TE-Model and TE-Info).

Note that we give only a brief overview of the HE-2 study here because it can only be addressed in the contexts of a longitudinal study, where participants revisit the exploration they previously completed. This area is the most uncertain as to the benefits, but also the validity of such a study is most at risk to confounding variables from other research questions. If we were to study this process by itself, we suspect the results would be influenced by the study design setup, and hence believe it is imperative to first understand the effect of earlier modeling and analysis phases.

F. Potential Threats to Validity

While not all threats to validity for this research can be known at this time, we explore some of them [26].

Conclusion Validity. These studies are at risk of the *reliability of measurement* threat, and we hope to form a community of reviewers for improved measurement of utility in this protocol. When studies are completed between-subjects, an argument must be made, by the individual study authors, to mitigate the *random heterogeneity of subjects* threat. *Low sample size* may threaten the conclusions of the experiments in Sec. III-C.

Construct Validity. Whenever combinations of construction, analysis, and evolution are evaluated with the same subjects,

there is an *interaction of different treatments* effect. This threat can be mitigated through replications with different combinations of treatment protocols. The *experimenter expectancies* threat should be mitigated by using explicit question wording in protocols whenever possible and documenting off-script questions that may bias subjects.

Internal Validity. We must take care to mitigate the *selection*, *compensatory rivalry* and *resentful demoralization* threats by adequately analyzing subjects' prior knowledge and views of GORE approaches prior to the study as well as evaluating their motivations for participating in the study. Proponents and opponents of goal modeling should be avoided in such studies.

External Validity. This line of research is motivated by external validity issues in past GORE studies. Each individual study will have the *interaction of selection and treatment* and *interaction of setting and treatment* threats, but it is our goal that all studies taken together can mitigate these threats by using different populations and settings, forming a body of knowledge about the utility of GORE approaches. To mitigate these threats, each study should document their context of observation (i.e., population demographics, model domain and size, as well as the specific GORE approach/languages used).

These studies were crafted from the lens of the iStar family of GORE approaches. We believe that at least the modeling portion of this methodology directly applies to other GORE approaches, such as KAOS, and that the study as a whole can be easily adapted for other approaches.

IV. Summary and Conclusions

Inspired by this year's theme, Crossing Boundaries and Increasing Impact, we reflect on some of the underlying assumptions made in the GORE literature and how these assumptions may have contributed to the lack of adoption of GORE approaches in industry. In this short communication, we divided GORE activities into the categories of construction, analysis, and evolution to generate assumptions that might interact with understanding of the utility of GORE approaches, and connected these with testable hypotheses and treatments. We contributed a study protocol consisting of a series of studies to evaluate where modelers find benefits in GORE approaches and explore opportunities for process automation.

At this stage, we focus on obtaining feedback on the studies themselves and to actively solicit collaborators in the RE community. We hope this work will initiate a discussion about how to evaluate the construct of utility, and discussions on how to increase the adoption of GORE approaches. We welcome involvement from others to ensure that the experiments proposed in this paper can be replicated across other goal modeling languages and tools. This protocol and supplemental information is available as a public gitHub project at *https://github.com/amgrubb/gore-study*, and we invite others to contribute to the project and document their protocols/studies prior to completing them. This will help us collect both positive and negative results. As in [15], this paper has become an instrument to consolidate our future work.

Acknowledgments. We would like to thank Marsha Chechik and our anonymous referees for helping improve this work.

References

[1] L. Chung, B. A. Nixon, E. Yu, and J. Mylopoulos, *Non-Functional Requirements in Software Engineering*. Kluwer Academic Publishers, Norwell, MA, 2000.

[2] E. Yu, "Towards Modeling and Reasoning Support for Early-Phase Requirements Engineering," in *Proc. of RE'97*, 1997, pp. 226–235.

[3] D. Amyot, "Introduction to the User Requirements Notation: Learning by Example," *Comput. Netw.*, vol. 42, no. 3, pp. 285–301, Jun. 2003.

[4] 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.

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

[6] C. Cares, X. Franch, A. Perini, and A. Susi, "Towards Interoperability of I* Models Using iStarML," *Comput. Stand. Interfaces*, vol. 33, no. 1, pp. 69–79, Jan. 2011.

[7] A. Anton and C. Potts, "The Use of Goals to Surface Requirements for Evolving Systems," in *Proc. of ICSE'98*, 1998, pp. 157–166.

[8] J. Horkoff and E. Yu, "Comparison and Evaluation of Goal-Oriented Satisfaction Analysis Techniques," *Requir. Eng.*, vol. 18, no. 3, pp. 199–222, 2013.

[9] F. Dalpiaz, A. Borgida, J. Horkoff, and J. Mylopoulos, "Runtime Goal Models: Keynote," in *Proc. of RCIS'13*, 2013, pp. 1–11.

[10] A. M. Grubb and M. Chechik, "Looking into the Crystal Ball: Requirements Evolution over Time," in *Proc. of RE'16*, 2016, pp. 86–95.

[11] ——, "Modeling and reasoning with changing intentions: An experiment," in *Proc. of RE'17*, 2017, pp. 164–173.

[12] Aprajita, S. Luthra, and G. Mussbacher, "Specifying Evolving Requirements Models with TimedURN," in *Proc. of MiSE'17*, 2017, pp. 26–32.

[13] S. Luthra, Aprajita, and G. Mussbacher, "Visualizing Evolving Requirements Models with TimedURN," in *Proc. of MiSE'18*, 2018.

[14] D. Lo, N. Nagappan, and T. Zimmermann, "How Practitioners Perceive the Relevance of Software Engineering Research," in *Proc. of FSE'15*, 2015, pp. 415–425.

[15] X. Franch, D. M. Fernández, M. Oriol, A. Vogelsang, R. Heldal, E. Knauss, G. H. Travassos, J. C. Carver, O. Dieste, and T. Zimmermann, "How do Practitioners Perceive the Relevance of Requirements Engineering Research? An Ongoing Study," in *In Proc. of RE'17*, 2017, pp. 382–387.

[16] A. Mavin, P. Wilkinson, S. Teufl, H. Femmer, J. Eckhardt, and J. Mund, "Does Goal-Oriented Requirements Engineering Achieve Its Goal?" in *In Proc. of RE'17*, Sept 2017, pp. 174–183.

[17] 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.

[18] J. Horkoff, N. A. M. Maiden, and D. Asboth, "Creative goal modeling for innovative requirements," 2018. Submitted.

[19] K. Abad, W. Pérez, J. P. Carvallo, and X. Franch, "I* in Practice: Identifying Frequent Problems in Its Application," in *Proc. of the Symposium on Applied Computing*, 2017, pp. 1122–1129.

[20] OED Online, *"utility, n."*, Oxford University Press, January 2018, www.oed.com/view/Entry/220771. Accessed 13 March 2018.

[21] "i* Wiki: i*-related PhD Dissertations," http://istar.rwth-aachen.de/tiki-index.php?page=i%2A-related+Phd+Dissertations, 2018, accessed: 2018-03-14.

[22] J. Horkoff and E. Yu, "Interactive Goal Model Analysis For Early Requirements Engineering," *Requir. Eng.*, vol. 21, no. 1, pp. 29–61, 2016.

[23] X. Franch, "A Method for the Definition of Metrics over i* Models," in *Proc. of CAiSE'09*, 2009, pp. 201–215.

[24] G. Mathew, T. Menzies, N. Ernst, and J. Klein, "Shorter Reasoning About Larger Requirements Models," in *Proc. of RE'17*, 2017.

[25] M. Bella and B. Hanington, *Universal Methods of Design*. Rockport Publishers, 2012.

[26] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*. Springer Berlin Heidelberg, 2012.