Ontology-based Modeling and Analysis of Trustworthiness Requirements: Preliminary Results

Glenda Amaral[1][0000−0003−0460−2271], Renata Guizzardi[2][0000−0002−5804−5741], Giancarlo Guizzardi[1][0000−0002−3452−553X], and John Mylopoulos[3][0000−0002−8698−3292]

[1] CORE/KRDB, Free University of Bozen-Bolzano, Bolzano, Italy
{gmouraamaral,giancarlo.guizzardi}@unibz.it
[2] NEMO/UFES, Espírito Santo (UFES), Brazil
rguizzardi@inf.ufes.br
[3] University of Ottawa, Ottawa, Canada
jm@cs.toronto.edu

Abstract. The advent of Artificial Intelligence (AI) technologies has made it possible to build systems that diagnose a patient, decide on a loan application, drive a car, or kill an adversary in combat. Such systems signal a new era where software-intensive systems perform tasks that were performed in the past only by humans because they require judgement that only humans possess. However, such systems need to be trusted by their users, in the same way that a lawyer, medical doctor, driver or soldier is trusted in performing the tasks she is trained for. This creates the need for a new class of requirements, *Trustworthiness Requirements*, that we have to study in order to develop techniques for their elicitation, analysis and operationalization. In this paper, we propose a foundation to develop such techniques. Our work is based on an *Ontology of Trust* that answers questions about the nature of trust and the factors that influence it. Based on the answers, we characterize the class of trustworthiness requirements. Among other things, this characterization supports the requirements engineer in defining thurstworthiness requirements, identifying the risks presented by the system-to-be, and understanding the signals the system must emit to gain and maintain trust.

Keywords: Trustworthiness Requirements · AI Systems · OntoUML.

1 Introduction

Trust is an essential ingredient of everyday life. We relate to people, organizations and things because we trust them to deliver on a certain goal, task or asset. Trust is especially important in the case of safety-critical services that can directly affect human lives, such as medical diagnosis, autonomous driving, military technology, terrorism detection, and other situations that pose risks to

human life and health. And although we tend to be tolerant if a"translation service produces grammatically incorrect sentences or if a cell phone camera misses to recognize a person" [12], tolerating the possibility of a single wrong decision in "critical decision-making systems such as security, healthcare, or finance, where human lives or significant assets are at stake" [12], is not acceptable. As systems are being developed, with or without AI technologies, that do make critical decisions, it is essential that their users trust them in the same way they trust their doctors, drivers and police. In the context of AI systems this was a key conclusion of the High-Level Expert Group on Artificial Intelligence (AI HLEG), which elaborated a set of ethics guidelines for trustworthy AI, as part of the European Strategy on Artificial Intelligence [9]. A similar conclusion was drawn in the "Explainable AI" initiative launched by the United States Defense Advanced Research Projects Agency (DARPA) [8], with the objective of making deep learning systems more trustworthy and controllable. These considerations call for studying a new class of requirements, namely, *Trustworthiness Requirements*, so that we can understand their nature and develop proper analysis techniques.

But what exactly is trust? And what makes a system trustworthy? In this work, we answer these questions in terms of a recently proposed Reference Ontology for Trust (ROT) [1]. Then, we combine ROT with the Non-Functional Requirements Ontology (NFRO) [7], which has the basic concepts to allow the definition of functional and non-functional requirements. This combination allows us to define the class of trustworthiness requirements and their relation to concepts such as trust, capability, vulnerability and risk, among others.

Here, we characterize trustworthiness requirements as a special class of *quality requirements* (in the sense of [7]) where the desired states-of-affairs are stakeholder mental states that include an *attitude* of trust towards the system-to-be. This trust is based on the system's track record in delivering its intended services (driving, diagnosing, decision-making, etc), the availability of valid information on that track record (no falsehoods or half-truths), as well as transparency on the delivery of the system's services.

The remainder of this paper is structured as follows. Section 2 discusses the ontological foundations in which our analysis is grounded. Section 3 introduces trustworthiness requirements and related concepts. Section 4 presents our proposal, a *Reference Ontology of Trustworthiness Requirements*. We conclude in Section 5 with some final considerations.

2 Research Baseline

In this paper, we provide an ontological analysis of trustworthiness requirements and trustworthiness-related risks, grounded in the Unified Foundational Ontology (UFO) [4]. In our analysis we shall rely on the trust-related concepts defined in the Reference Ontology of Trust proposed in [1] and on the ontological interpretation of non-functional requirements presented in [7].

The Reference Ontology of Trust (ROT). Based on UFO, Amaral et al. [1] present a Reference Ontology of Trust (ROT) which formally characterizes the concept of trust, as well as clarifies the relation between trust and risk, and represents how risk emerges from trust relations.

ROT makes the following ontological commitments on the nature of trust:
- Trust is always about a trustor's intention. An agent trusts another only relative to a goal, for the achievement of which she counts upon the trustee.
- Trust is a complex mental state of a trustor regarding a trustee. This complex mental state is composed of: (i) a trustor's intention, whose propositional content is a goal of the trustor; (ii) the belief that the trustee has the capability to perform the desired action; and (iii) the belief that the trustee's vulnerabilities will not prevent her from performing the desired action.
- The trustor is necessarily an "intentional entity". Briefly put, the trustor is a cognitive agent, an agent endowed with goals and beliefs.
- The trustee is not necessarily a cognitive system. The trustee is an entity capable of having an impact on a goal of the trustor.
- Trust is context dependent. The trustor may trust in the trustee in a certain context but may not trust her for the same action in a different context.
- Trust always implies risk. By trusting, the trustor accepts to become vulnerable to the trustee in terms of potential failure of the expected action and result, as the trustee may not perform the expected action or the action may not have the desired result.

Figure 1 depicts a ROT excerpt, which captures most of the aforementioned ontological notions. As in the original ROT ontology, this model is represented in OntoUML [5]

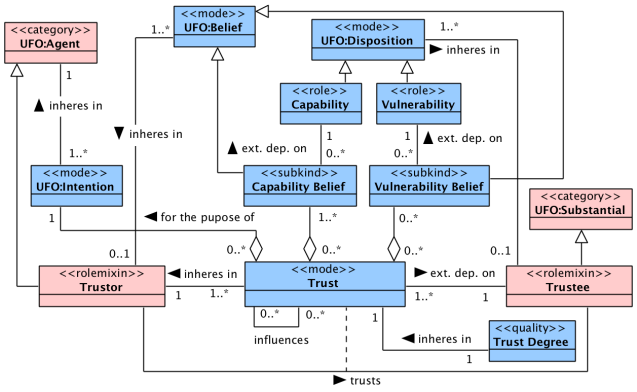

Fig. 1. A fragment of ROT depicting the mental aspects of trust

In ROT, Trust is modeled as a complex mode (a dependent entity) composed of a Trustor Intention, whose propositional content is a goal of the Trustor, and a set of Beliefs that inhere in the Trustor and are exter-

nally dependent on the *dispositions* [2, 6] that inhere in the Trustee. These beliefs include: (i) the Belief that the Trustee has the Capability to perform the desired action (Capability Belief); and (ii) the belief that the Trustee's Vulnerabilities will not prevent her from performing the desired action (Vulnerability Belief). The Trustee's Vulnerabilities and Capabilities are dispositions that inhere in the Trustee, which are manifested in particular situations, through the occurrence of events [6].

ROT relies on the Common Ontology of Value and Risk (COVER) proposed by Sales et al. [15] to represent the relation between trust and risk. A central notion for characterizing risk in COVER is a chain of events that impacts on an agent's goals, which the authors name Risk Experience. Risk Experiences focus on unwanted events that have the potential of causing losses and are composed by threat and loss events. A Threat Event is the one with the potential of causing a loss, which might be intentional or unintentional. A Threat Event might be the manifestation of a Vulnerability (a special type of disposition whose manifestation constitutes a loss or can potentially cause a loss from the perspective of a stakeholder). The second mandatory component of a Risk Experience is a Loss Event, which necessarily impacts intentions in a negative way [15]. When actions related to a trust relation are performed, they may satisfy the goals of the trustor or, in the worst case, they may not have the desired result. In this case, the resulting situation stands for a Threat Situation that may trigger a Threat Event, which may cause a loss. The Loss Event is a Risk Event that impacts intentions in a negative way.

The Ontology of Non-functional Requirements (NFRO). In [7], the authors propose a UFO-based ontological interpretation of non-functional requirements. In NFRO, requirement is defined as a goal. Requirements are specialized into NFRs (also named quality goals) and functional requirements (FRs). FRs refer to a function (a capability, capacity) that has the potential to manifest certain behavior in particular situations, while NFRs refer to qualities taking quality values in particular quality regions. For example, a software system is considered to have good usability if the value associated to its "usability" requirement maps to a region "good" in the "usability" quality space. Figure 2 depicts a selected subset of the NFRO that are relevant for our discussions on trustworthiness requirements. For an in-depth discussion and formal characterization of qualities, quality types, quality regions, and quality spaces, refer to [4].

3 Trustworthiness Requirements

Requirements are prescriptions of intended states-of-affairs that the system-to-be should bring about. Traditionally, these states-of-affairs were system-related, such as functions the system should deliver, or qualities it should possess with respect to performance, reliability, usability etc. Social requirements and physical requirements have been introduced in the literature more recently with the advent of socio-technical and cyber-physical systems [11, 13]. For example, "schedule meeting" is a social requirement because the desired state-of-affairs is one

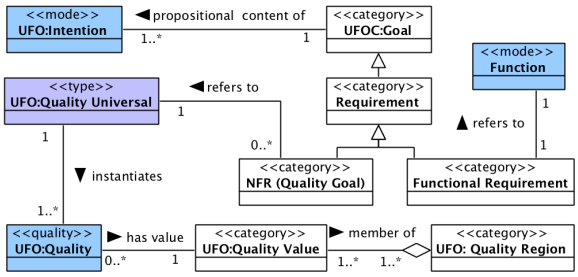

Fig. 2. A fragment of the Ontology of Non-functional Requirements

that includes a new meeting, where meeting is a social artifact (a bundle of rights, commitments, powers, etc. binding a number of participants). On the other hand, "distance from nearby physical objects $\geq$ 50cm" is a physical requirements for an autonomous vehicle. Personal requirements constitute a forth category of requirements where the desired states-of-affairs involve attitudinal (mental) properties of (some of) the system's stakeholders. For example, "$\geq$70% of departments members are using the meeting scheduling system" is a personal requirement (more specifically, an acceptance requirement) in that the system-to-be has to bring members of the department to a state of mind where they are willing to use the system. Trustworthiness requirements are personal requirements as well in that their desired states-of-affairs are ones where some of the stakeholders trust the system.

But how can an agent earn the trust of the recipients of its services? Firstly, the agent can make available to its users its credentials (degrees, accreditations, certificates, awards) that suggest that "it knows its craft", "it is doing a good job", and the like. The agent can also make available information on its track record, such as reviews from service recipients and statistics on its experience. Moreover, all information that is used must be true (no half-truths and no lies). Politicians are able to convince a certain segment of their electorate to trust them; However, if done through the use of half-truths and lies in the process, this can make them trusted but unworthy of trust, or untrustworthy.

Trustworthiness means more than trust in other ways as well. A trustworthy agent must be delivering its service in a professional and effective manner. For example, a medical doctor agent may be trusted by most of its patients because of its accreditations and its affiliation with a healthcare organization, but it is not trustworthy unless it also delivers reliable healthcare services to its patients. Reliability here includes availability, the good doctor is available when you need it, but also effectiveness in its diagnoses and treatments of its patients. A medical doctor you rarely succeed to make an appointment with isn't trustworthy, nor is one whose diagnoses are often contradicted by expert colleagues.

Another element of trustworthiness is transparency in the delivery of an agent's services. Transparency is influenced by many factors [10]. In the context

of an agent delivering a service, transparency includes offering information on what the agent is doing, as well as rationale for its decisions (aka explainability).

On the basis of these considerations, a trustworthiness requirement can be AND-refined into a *reliability requirement*, a *truthful information communication requirement* and a *transparency requirement* for the service being delivered.

Trustworthiness requirements are quality requirements [7]. This means that they constrain the level of presence of a quality in its subject. For example "being red" is a property that constrains the colour quality of its subject to be in the red region of a color quality space (a chromatic map known as the color spindle). Likewise "being trustworthy" is a constraint for agents or services to fall in the trustworthiness region of a space that also includes an untrustworthiness region.

Of course, trustworthiness isn't only a black-and-white quality requirement. It also includes weaker versions that can be defined by refinement operators [7]:

- Probabilistic refinements: These consider what percentage of the uses of the system's services were deemed trustworthy by the recipients of these services. For example, for a diagnostic system, a trustworthiness requirement could be "≥80% of uses were found trustworthy";
- Fuzziness refinements: Here, we weaken the notion of trustworthiness by making it fuzzy to include things that are "almost trustworthy","fairly trustworthy", "definitely not untrustworthy".
- Subjectivity refinements: These are requirements of the form "≥N% of users asked consider the system trustworthy". Note that unlike probabilistic refinements, subjectivity refinements focus on users, not uses.

These refinement operators can also be applied to the sub-goals of a trustworthiness requirement, to yield a full space of requirements concerning the trustworthiness quality.

4 Ontology-based Modeling and Analysis of Trustworthiness Requirements

Understanding the elements of stakeholder trustworthiness towards the system to be is important because they reveal the qualities and properties the system should have in order to be considered trustworthy and effectively promote well-placed trust. Note that as trust is contextually dependent (the trust degree of a trustor in a trustee may vary from a context to another) the implementation of trustworthiness requirements depends on the specific application. For example, a user trusts a system in collecting her location data but not when she is in sensitive places, such as when she is being treated at a hospital, since such information may lead to disclose a health issue.

Another advantage of making the components of trust explicit is that this knowledge can be used as input to the definition of trust-warranting signals that ensure trustworthy behavior. In other words, once the system's capabilities and vulnerabilities related to the trust of the stakeholder are known, it is possible to reason about the signals that the system should emit to indicate that it is

capable of successfully realizing the capabilities and prevent the manifestation of the vulnerabilities. For example, information about how privacy and security measures are implemented could be provided as signals of the trustworthiness of a system. Other relevant examples of trust-warranting signals are data certificates and data provenance information, both relevant for systems dealing with large amounts of data, to avoid bias and unfair results.

Finally, the identification of trust components is equally important to the assessment of risks related to the capabilities and vulnerabilities, which are the focus of stakeholders' beliefs. As previously discussed, *capabilities* are dispositions that inhere in an agent and, as such, are manifested in particular situations, through the occurrence of events [6]. As defined in the Common Ontology of Value and Risk (COVER) [15], a *threat event* is a type of *risk event* that may be the manifestation of a capability of the system, in case it fails to realize this specific capability in order to bring about an outcome desired by the stakeholder. According to COVER, the threat event may lead to a *loss event*, which negatively influences the stakeholder's intention. For example, suppose that a network malfunction prevents a medical system to access the server containing patient data and, as a result of that, it cannot deliver its capability of providing a diagnosis. In this case, the network malfunction is a threat event, which leads to a lack of diagnosis loss event.

Similarly, *vulnerabilities* are also a special type of disposition, whose manifestation causes or can potentially cause a loss, under the perspective of a stakeholder. Therefore, a threat event may be the manifestation of a vulnerability and eventually trigger a loss event. To illustrate this point, let us imagine that our medical system has a security vulnerability and is thus hacked, leading to the leak of patient data. In this case, the hacking threat event, resulting from the manifestation of the system's security vulnerability, triggered the patient privacy loss event.

We represent the concepts related to Trustworthiness Requirements in the OntoUML model depicted in Fig. 3, and the emergence of risks in this scenario in Fig. 4.

As shown in Fig. 3, we modeled Requirement as a Goal, which is the propositional content of an Intention of a Stakeholder. Quality Requirement is a type of Requirement, and Trustworthy Requirement is a type of Quality Requirement. Stakeholders are represented as Agents that play the role of trustor, while the System is an existentially independent object that plays the role of trustee. The System intends to satisfy the Trustworthiness Requirements.

As pointed out in section 3, the analysis of the trustworthiness requirement involve its decomposition in three other quality requirements, namely, *reliability requirement*, *truthful information communication requirement* and *transparency requirement*. Thus, we include in the model of Fig. 3, a composition relation between Trustworthiness Requirement and Quality Requirement. Additionally, this model supports the representation of the mentioned sub-requirements as instances of the Quality Requirement concept. All Qual-

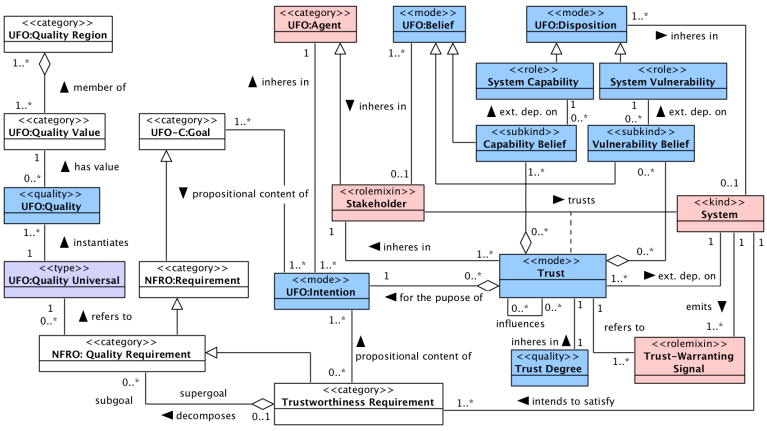

Fig. 3. Modeling trustworthiness requirements in OntoUML

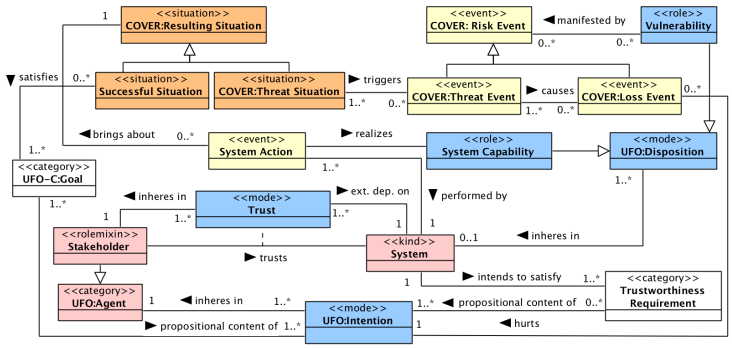

Fig. 4. Modeling the emergence of trustworthiness-related risks

ity Requirements are such that they restrict the value of the qualities at hand to a particular set of values of the corresponding Quality Region. Trustworthy Requirement restrict the values of qualities referring to reliability, transparency and information truthfulness to particular set of values accordingly.

As for Trust, we represent it as a complex mode composed of a Stakeholder's Intention, whose propositional content is a Goal of the Stakeholder, and a set of Beliefs that inhere in the Stakeholder and are externally dependent on the dispositions [2, 6] that inhere in the System. These beliefs include: (i) the Belief that the System has the Capability to perform the desired action (Capability Belief); and (ii) the Belief that the System's Vulnerabilities will not prevent it from exhibiting the desired behavior (Vulnerability Belief). The System's Vulnerabilities and Capabilities are dispositions that inhere in the System, which are manifested in particular situations, through the occurrence of events [6]. We adopt the interpretation of capability proposed by Azevedo et al. [2], who defined capability

as the power to bring about a desired outcome. As previously discussed, the System can emit Trust-Warranting Signals in order to indicate that it is capable of successfully realizing the capabilities and prevent the manifestation of the vulnerabilities.

All these ontological concepts play an important role in helping us understand if the system is compliant to the *reliability requirement*, *truthful information communication requirement* and *transparency requirement*, composing the trustworthiness requirement. For example, for *reliability*, we must understand how much of the Stakeholder's Capability Belief is actually met by the results of the system's operation (i.e., by system actions); regarding *truthful information communication*, the System Capability of providing truthful information may be validated, by comparing the information generated by the system with information known to be real; and finally, regarding *transparency*, we must make sure that the Trust-Warranting Signals are enough to make the Stakeholder satisfied w.r.t how often and how well the system explains its decision-making process.

We represent the emergence of trustworthiness-related risks in the OntoUML model depicted in Fig. 4. In order to realize the Capabilities, the System performs some Actions that bring about a Resulting Situation. The Resulting Situation may satisfy the Stakeholder's Goals (and in this case it is considered a Successful Situation) or, in the worst case, it may not have the desired result and the Stakeholder will not be able to achieve her goal. In this case, the Resulting Situation stands for a Threat Situation that may trigger a Threat Event, which may lead to a Loss Event that impacts intentions in a negative way, as it hurts the Stakeholder's Intentions of reaching a Goal. Analogously, System's Vulnerabilities may enable the occurrence of Risk Events that, in the worst case, may cause a Loss Event which will hurt the Stakeholder's Intentions of reaching her Goal.

5 Final Remarks

In this paper, we presented an ontological analysis characterizing the concept of trustworthiness requirements of software systems, as well as the emergence of risks when using such system.

The elicitation of trust requirements has been broadly studied and different approaches have been proposed in the literature to support the capture and implementation of trust requirements in the context of software systems [3,9,11,14]. Despite the wide number of efforts to properly analyse trustworthiness requirements and trust-related issues, little has been said about what constitutes the stakeholders' trust in the system, what it depends upon and how trustworthiness-related risks can be identified. Differently from other approaches, our proposal analyses the components of the trust complex mental state of the trustor in order to identify what the system should have for stakeholders to trust it. These elements are fundamental for a better understanding and proper elucidation of trustworthy requirements. Moreover, they are key for the identification of

trustworthiness-related risks that may arise when the requirements are not fulfilled accordingly.

As future work, we plan to further validate our ontology by doing real case studies and having experts evaluate the results. We also plan to define ontological patterns, based on this ontology, to support the modeling and analysis of trustworthiness requirements, aiming at facilitating the development of trustworthy systems. Finally, we plan to propose a systematic process for identifying trustworthiness requirements, grounded on these patterns and on the ontological account of trustworthiness requirements presented here.

Acknowledgment

CAPES (PhD grant# 88881.173022/2018-01) and OCEAN project (UNIBZ).

References

1. Amaral, G., Sales, T.P., Guizzardi, G., Porello, D.: Towards a Reference Ontology of Trust. In: Proc. of CoopIS. pp. 3–21. Springer (2019)
2. Azevedo, C.L.B. et al.: Modeling resources and capabilities in enterprise architecture: A well-founded ontology-based proposal for ArchiMate. Information systems **54**, 235–262 (2015)
3. Giorgini, P. et al.: Modeling social and individual trust in requirements engineering methodologies. In: Intl. Conf. Trust Management. Springer (2005)
4. Guizzardi, G.: Ontological foundations for structural conceptual models. Telematica Instituut / CTIT (2005)
5. Guizzardi, G., Wagner, G., Almeida, J.P.A., Guizzardi, R.S.S.: Towards ontological foundations for conceptual modeling: the Unified Foundational Ontology (UFO) story. Applied ontology **10**(3-4), 259–271 (2015)
6. Guizzardi, G. et al.: Towards ontological foundations for the conceptual modeling of events. In: Proc. of 32nd ER. pp. 327–341. Springer (2013)
7. Guizzardi, R. et al.: An ontological interpretation of non-functional requirements. In: Proc. of FOIS. vol. 14, pp. 344–357 (2014)
8. Gunning, D., Aha, D.W.: Darpa's explainable artificial intelligence program. AI Magazine **40**(2), 44–58 (2019)
9. Hleg, A.I.: Ethics Guidelines for Trustworthy AI. B-1049 Brussels (2019)
10. Leite, L., Cappelli, C.: Software transparency. Bus Inf Syst Eng **2**(3) (2010)
11. Mohammadi, G.: Trustworthy Cyber-Physical Systems. Springer (2019)
12. Nassar, M., Salah, K., ur Rehman, M.H., Svetinovic, D.: Blockchain for explainable and trustworthy artificial intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **10**(1), e1340 (2020)
13. Paja, E., Chopra, A.K., Giorgini, P.: Trust-based specification of sociotechnical systems. Data & Knowledge Engineering **87**, 339–353 (2013)
14. Rosemann, M.: Trust-aware process design. In: International Conference on Business Process Management. pp. 305–321. Springer (2019)
15. Sales, T. et al.: The Common Ontology of Value and Risk. In: Proc. of 37th International Conference on Conceptual Modeling (ER)