

Towards the Concept of Trust Assurance Case

Emilia Cioroica*, Barbora Buhnova†, Daniel Schneider*, Emrah Tomur‡, Ioannis Sorokos* and Thomas Kuhn*

* *Fraunhofer IESE*, Kaiserslautern, Germany

† *Masaryk University*, Brno, Czech Republic

‡ *Ericsson*, Istanbul, Turkey

*{emilia.cioroica, ioannis.sorokos, daniel.schneider, thomas.kuhn}@iese.fraunhofer.de †{buhnova@fi.muni.cz}

Abstract—Trust is a fundamental aspect in enabling a smooth adoption of robotic technical innovations in our societies. While Artificial Intelligence (AI) is capable to uplift digital contributions to our societies while protecting environmental resources, its ethical and technical trust dimensions bring significant challenges for a sustainable evolution of robotic systems. Inspired by the safety assurance case, in this paper we introduce the concept of trust assurance case together with the implementation of its ethical and technical principles directed towards assuring a trustworthy sustainable evolution of AI-enabled robotic systems.

Index Terms—Trust, Safety, Ethics, Dependability, Technical Trust, Human Trust

I. INTRODUCTION

Emerging robotic systems, under the control of AI components, are requested to react to changes in their environments, such as changes imposed by the use of dedicated driving paths. In this new technological landscape, the process of assuring trust is challenging, with multiple considerations.

From the technical perspective, the non-deterministic nature of AI components operating in open contexts yields a significant threat: hard-to-detect malicious behavior can be hidden in the control of a robot. Aiming at improving itself during operation, an AI component, over time, is permitted to provide a different set of output values for one single set of inputs. While this freedom increases versatility and usefulness of an intelligent robot, it also creates perfect conditions for triggering malicious behavior.

On top of the technical concerns, AI deployment within the robotics domain is raising ethical concerns as well. Besides concerns on the role human beings will play in the emerging AI-socio-technological ecosystems formed around intelligent robots, the subjective and context-dependant nature of trust contain affect and a moral dimensions hard to formalize in general models of trust [1].

While context-specific solutions designed for gaining the human affect [2] are hard to leverage between domains, the contradicting moral aspects of trust are even more far away from being solved. Human morality falls in either the deontological or the consequentialist landscape. Deploying robotic systems that help in decision-making within a safety-critical domain (like transportation) is challenging. While a reduction in the number of deaths is a strong argument from the consequentialism viewpoint, from the deontological perspective this argument is unacceptable. In a deontological landscape, it is hard to justify a technical intervention that

occasionally causes the loss of peoples' lives by arguing that society becomes safer overall. On top of this, the psychological trait of humans shows that despite a statistical number of deaths decreases, the numbers are assigned human figures and stories to which people relate to [3].

In general, the design of robotic systems, capable to consider trust in its full dimensions, is challenged by its highly contextual and subjective nature that continuously develops with new and emerging understandings. In the domain of robotics in particular, it is even more difficult to design trust systems that work with trust in their full technical and human dimensions to achieve decisions that impact societies on a large scale.

Moving from the current trends of designing trust models and trust algorithms [4], which become obsolete or impossible to leverage with every new advancement of trust understandings, in this paper, we instead propose the concept of a *Trust Assurance Case*, given as a set of actionable principles that account of trusts' multiple dimensions for assuring its existence via guiding the co-evolution of human and technical trust for robotic systems.

To this end, Section II presents human and technical trust concerns that are used in conceptualizing the Trust Assurance Case introduced in Section III. Section IV then presents our strategy for implementing the trust assurance case in a one-to-one mapping to its core principles.

II. ASPECTS OF TRUST

A. From human to digital trust

Trust is considered both a belief-based and computation-based concept [5], being human, social or system-centered. Even though characterized by subjectivity, one aspect is certain: in our societies, trust is implicit. We notice it as we notice air, only when it becomes scarce or polluted.

For an individual entering a society, the loss or pain of losing trust becomes greater than the reward of gaining it [6]. But even though the costs out-weight the benefits, a human being is still willing to give his/her trust. Humans decide to put themselves into the hands of entities they do not fully know or understand, based on the belief that those entities can be removed from power [7].

This means that AI-controlled robots are also capable of gaining trust in our societies. People can trust to put areas of their lives under the control of AI-based robots, based on the belief that this control can be removed in case it fails to fulfill

expectations. For example, an individual can decide to trust an autonomous robot controlled by an AI component, if another entity that is highly trusted can take over control in case of detected deviations and unmet expectations.

B. Safety Assurance

When humans think about the trust they are mainly concerned with their safety and the safety of their environment. Secondly, people are concerned about security threats. An investigation of trust concerns, therefore, requires the foremost consideration of safety aspects. In the safety domain, assurance cases [8] have long been used to increase knowledge by making the strength of arguments explicit. The safety assurance case, as an instrument for gaining the trust of certifying authorities, enables stakeholders with various skills to reason about the safety of a system. Even though safety cases are centered around systems only, and operation in a fixed context, the creation of trust assurance cases can be inspired by the safety philosophy as we detail in Subsection III-A. That is because first and foremost, evidence-based knowledge is capable to remove the fear of the unknown. Second, the subjective nature of the safety assurance process is adjusted to the human nature of trust already. Third, the rigorous nature of a safety case can systematically guide the engineering of trusted systems and it needs to be uplifted to the demands of dynamic and complex trust concerns.

C. Pinpointing the object of trust

The unit of trust evaluation is the observed goals. A goal is an evidence of accepted objective fulfilled by system agents [9]. Therefore, in conceptualizing trust assurance cases, the observed goal is the vehicle that transports evidence from lower computational levels to the upper levels of ethical and strategic decisions. Further on, runtime goal evaluation is the mechanism capable to account entities for their actions, responses, achievements, and undesired behaviors. And in scenarios where AI components evolve at runtime, mechanisms for predicting their goals at runtime need to be considered as well. Only based on observed predicted behavior a supervising entity can decide where to place the control of the system: to the complex AI component or to a much simpler, highly understood, and safe proven fail-over behavior.

III. THE CONCEPT OF TRUST ASSURANCE CASE

We envision the Trust Assurance Cases that address both the human and the technical aspects of trust through a multi-layer framing concept that enables dynamic risk assessment based on runtime prediction of goals. The trust assurance case can be used for supporting diligent engineering and holistic quality assurance of emerging intelligent robots.

A. Methodology

We define the concept of Trust Assurance Case by uplifting the principles of well-established Safety Assurance Case [8] and trust concerns gathered from literature and summarized in Section II. In conceptualizing the trust assurance case, the

static nature of the safety assurance needs to be adapted to the dynamic nature of trust evaluation. While the safety assurance case is a static tool that works with textual description, the trust assurance case is envisioned to support the runtime dynamic evaluation of a robot's functional and non-functional behavior under the control of a goal-oriented AI component. This distinction is motivated by the dynamic adaptive nature of AI control and the object of trust evaluation which is the observed goal of AI evolution.

Relying on complex moral and affect argumentation, the subjective trust evaluation needs to be framed in contexts that are clearly communicated to a human evaluator. This is similar to the goal structuring [8] in safety assurance cases, but the source of evidence is different. While safety assurance cases work with evidence provided by assurance artifacts collected during development, the trust evaluation is based on computing evidence derived at runtime in specific technical situations. Based on runtime evidence, the argumentation part is structured in evaluation scenes that support the complex moral and affect subjective evaluation of human observers.

Similar to the safety evaluation, the trust evaluation is context-based. Given the subjective nature of both concepts: safety and trust, evaluation of a system's behavior needs to be placed in a context. In the same way, any system can be safe or unsafe if used inappropriately, any system can be trusted or distrusted depending on its operational context.

Last, the main steps in the process of safety assurance presented in [8] have been uplifted to technical and moral concerns of trust. Existing evidence shows that visualization techniques [2] can be used to design effective and efficient communication with a human evaluator.

B. The Principles of Trust Assurance Case

In the vulnerable technological and moral settings of AI evolution, the benefit of the trust assurance case is the clear communication of arguments supported by computed evidence. The proposed process of assuring trust in AI-controlled robots, as depicted in Fig. 1 is based on multiple principles:

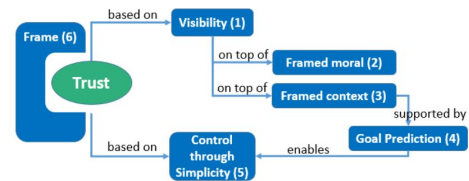


Fig. 1. Principles of Trust Assurance for AI-controlled systems

- 1) **Visibility of arguments for change in the behavior of AI-based robots.** The ultimate gain of trust is through human supervision. Humans are visual beings, therefore images are both efficient and effective in communicating information. In the process of assuring trust, arguments can effectively and efficiently be formed based on image representation of a robot's intended behavior in complex technical and social settings. Similarly, in the process of

safety assurance, the GSN (Goal Structuring Notation) is the mechanism that clearly communicates arguments [8].

- 2) **Framing the moral considerations.** Morality is diverse and can be judged only with framed considerations. Similar to the process of assuring safety, where the context framing makes the scope of evaluation specific, this principle makes the moral evaluation of intelligent robots specific while enabling the transfer of moral trust and reputation between societies.
- 3) **Framing the technical and context settings.** Trust outside the context tends to be wrongly considered the ultimate truth. When evaluated within a specified context, trust becomes both attainable and realistic as it enables argumentation for context change. When AI operates in a new technical setting, evidence of trusted behavior needs adjustments to the demands of the new setting. When the context changes, adjustments of computed reputation need to be done in accordance with the compatibility between the source and target community [10]. This principle adheres to the concept of framed context existing also in the safety assurance process.
- 4) **Goal prediction.** This principle enables evidence-based supervision of an external entity capable to decide on the course of action. The external entity can be a human being or a highly trusted monitoring mechanism that starts the fail-over behavior in case of detected deviations. In the process of building sustainable human trust, during the evolution of AI-controlled systems, both the moral and the technical control start under human supervision [11]. The technical control, then, is gradually replaced by highly trusted monitoring components. These monitoring components compare the predicted goals to the actual execution of the AI component and allow the AI component to control the robot only when it is in conformance with valid goals. This principle elevates the clear goal communication present in safety assurance by addressing the evolving nature of an AI component through behavior prediction.
- 5) **Control through simplicity.** The trust assurance process in AI-controlled robots requires the presence of a simpler, highly trusted mechanism ready to take over control in case the AI component fails to be trusted. This principle is motivated by the fact that rational entities can put their trust in control of complex entities if a fail-over safe behavior is ready to take over [7].
- 6) **Framing of trust acceptance criteria.** From the human perspective, reaching absolute trust is an unobtainable goal, as the truth itself has many facets. From the system perspective, clear acceptance criteria provide end goals for validity. This in turn enables the exchange of technology as it advances. Therefore this principle has the scope of specifying the "what" and enabling the exchange of the "how". On top of principle 2 and principle 3 this principle supports standardization activities. Trust assurance for safety critical systems needs

certification according to existing standards. Therefore, this principle is defined in the virtue of technological neutrality following the European directives and for keeping up with the fast technological progress.

IV. EVALUATION PLATFORM

We envision a strategy for implementing a simulation framework that is aimed to enable the dynamic execution of a trust assurance case according to the principles listed in the previous section. The framework supports the integration of modular components within the automotive domain [12] for integrating complex functions under evaluation.

A. Architecture

Fig. 2 depicts the main components of a framework that enables dynamic trust assurance of intelligent robots. At this stage, the concepts are technologically neutral in the sense that an architectural component can be implemented with any suitable technology. E.g., the AI component can be a Deep Learning algorithm or a Neural Network, another AI component or simulation technologies can be used for runtime prediction of goals.

The human trust evaluation needs to be framed into scenes. Each *Scene* aggregates multiple *Context Models* that describe the environment of the evaluation. Further on, the technical trust is evaluated based on the execution of *System Models*. Trust assurance of a system model that contains an *AI component* needs to have in its composition a *Monitor* that observes the behavior of the AI component and how well it is adhering to *Predicted Goals*. For the technical evaluation of goals, thresholds for allowed values can be defined, whereas for the human evaluation of trust user expectations and subjective moral judgments are considered. The Monitor informs a *Decision Control* Component of detected deviations and the decision control decides whether a highly trusted *Automated Control* should be triggered or not. To assuring the provision of technically trusted services and trusted behavior, the AI component remains in control of the system only if it complies with technical and user expectations. Results of the technical and moral evaluation are quantified in *Reputation* scores.

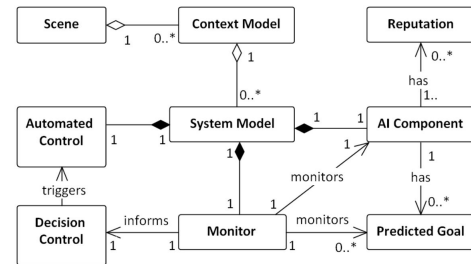


Fig. 2. Dynamic trust evaluation concept according to Trust Assurance Cases

B. Prototypical Implementation

In Fig. 3, an AI-controlled robot starts evolving under the supervision of a human observer. The principles of trust assur-

ance can be implemented following the architecture presented in Fig. 2 in the following way:

Implementation of Principle 1: The human is in charge of supervising both the moral decisions of the system and the technical execution of goals. While moral validation requires longer supervision, technical supervision is gradually replaced with automated monitored supervision. The gradual replacement of human technical supervision is depicted with a light-hashed pattern on the lower triangle. Visibility of arguments can be provided through the integration of a 3D Engines, for example Unity3D [13].

Implementation of Principle 2: The moral decisions of the AI-controlled robots can be evaluated in settings defined by the human observer in configuration engines that enable the user-friendly description of logical flows and specification of allowed intervals for deviations within accepted intervals. One example of such tool is Blockly [14].

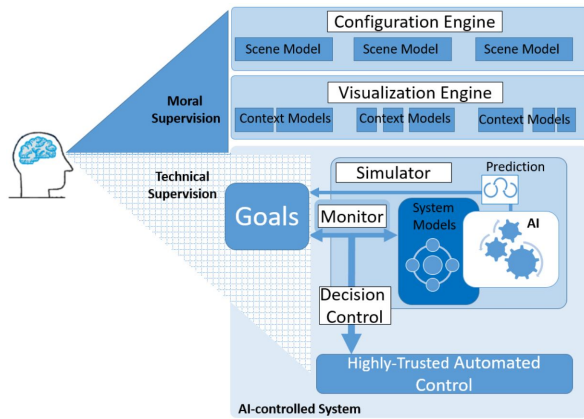


Fig. 3. Frameworks for dynamic evaluation of Trust Assurance Cases

Implementation of Principle 3: 3D engines enable the definition of objects and accurate representation of the real world in virtual environments. Linked to the moral context defined in the configuration engine, scenes that describe the technical settings can be loaded within 3D engines. For safety-critical systems under the control of AI components, information on technical context needs to be displayed to a human supervisor on a screen.

Implementation of Principle 4: For predicting goals, different technologies can be used [15]. The prediction can be performed through the execution of abstracted models of AI components. The models need to be executed in relation to models of the interacting components. The abstractions provide thresholds of valid values against which the AI execution is monitored and directed toward the scope of the evaluation.

Implementation of Principle 5: The trust evaluation starts under human supervision, thus the monitored data needs to be converted into information displayed in the 3D Engine. The human observer needs to have the possibility to trigger a safe fail-over behavior when the AI control is not trusted.

Implementation of Principle 6: is directed towards enabling the definition of trust interfaces. For the implementation

of this principle, dynamic contracts can be deployed [16]. Such contracts specify demands and guarantees between interconnected components, with those requesting the demands being able to offer corresponding guarantees with given levels of quality and trustworthiness. For enabling valid goal prediction based on the execution of abstract simulation models, demands and guarantees need to be specified in terms of value ranges.

V. CONCLUSIONS

In this paper, we introduced the concept of the Trust Assurance Case together with its underlying principles and conceptual implementation of a framework meant to support its execution. On the ideas and concepts listed in this paper, we will further investigate concepts as well as supporting means for achieving the stated goals in the trust assurance case.

Acknowledgment:

This work was supported by the project BIECO (www.bieco.org) which received funding from the European Union's Horizon 2020 research and by ERDF "CyberSecurity, CyberCrime and Critical Information Infrastructures Center of Excellence" (No. CZ.02.1.01/0.0/0.0/16_019/0000822).

REFERENCES

- [1] S. Marsh, T. Atele-Williams, A. Basu, N. Dwyer, P. R. Lewis, H. Miller-Bakewell, and J. Pitt, "Thinking about trust: People, process, and place," *Patterns*, vol. 1, no. 3, p. 100039, 2020.
- [2] R. Häußelschmid, M. von Buelow, B. Pfleging, and A. Butz, "Supporting trust in autonomous driving," in *Proceedings of the 22nd international conference on intelligent user interfaces*, 2017, pp. 319–329.
- [3] D. Ariely, "Seeing sets: Representation by statistical properties," *Psychological science*, vol. 12, no. 2, pp. 157–162, 2001.
- [4] S. A. Siddiqui, A. Mahmood, Q. Z. Sheng, H. Suzuki, and W. Ni, "A survey of trust management in the internet of vehicles," *Electronics*, vol. 10, no. 18, p. 2223, 2021.
- [5] B. Qureshi, G. Min, and D. Kouvatso, "Collusion detection and prevention with fire+ trust and reputation model," in *2010 10th IEEE International Conference on Computer and Information Technology*. IEEE, 2010, pp. 2548–2555.
- [6] A. Baier, "Trust and antitrust," *ethics*, vol. 96, no. 2, pp. 231–260, 1986.
- [7] P. Dasgupta, "A matter of trust: Social capital and economic development," *ABCDE*, vol. 119, 2011.
- [8] T. Kelly and R. Weaver, "The goal structuring notation—a safety argument notation," in *Proceedings of the dependable systems and networks 2004 workshop on assurance cases*. Citeseer, 2004, p. 6.
- [9] A. Cailliau and A. van Lamsweerde, "Assessing requirements-related risks through probabilistic goals and obstacles," *Requirements Engineering*, vol. 18, no. 2, pp. 129–146, 2013.
- [10] S. Ruohomaa and L. Kutvonen, "Trust management survey," in *International Conference on Trust Management*. Springer, 2005, pp. 77–92.
- [11] E. Cioroica, B. Buhnova, T. Kuhn, and D. Schneider, "Building trust in the untrustable," in *2020 IEEE/ACM 42nd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. IEEE, 2020, pp. 21–24.
- [12] P. Feth, T. Bauer, and T. Kuhn, "Virtual validation of cyber physical systems," in *Software-engineering and management 2015*, U. Abmann, B. Demuth, T. Spitta, G. Püschel, and R. Kaiser, Eds. Bonn: Gesellschaft für Informatik e.V., 2015, pp. 201–206.
- [13] "Unity 3d in the automotive domain." [Online]. Available: <https://unity.com/solutions/automotive-transportation-manufacturing>
- [14] "Google blockly," [Online]. At: <https://developers.google.com/blockly>
- [15] D. Misra, A. Bennett, V. Blukis, E. Niklasson, M. Shatkhin, and Y. Artzi, "Mapping instructions to actions in 3d environments with visual goal prediction," *arXiv preprint arXiv:1809.00786*, 2018.
- [16] D. Schneider and M. Trapp, "Conditional safety certification of open adaptive systems," *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, vol. 8, no. 2, pp. 1–20, 2013.