Bridging Trust in Runtime Open Evaluation Scenarios

Emilia Cioroaica*, Barbora Buhnova†, Eda Marchetti ‡, Daniel Schneider*, and Thomas Kuhn*
* *Fraunhofer IESE*, Kaiserslautern, Germany
† *Masaryk University*, Brno, Czech Republic
‡ *ISTI-CNR*, Pisa, Italy
*{emilia.cioroaica, thomas.kuhn, daniel.schneider}@iese.fraunhofer.de
†{buhnova@mail.muni.cz} ‡{eda.marchetti@isti.cnr.it}

***Abstract*—Solutions to specific challenges within software engineering activities can greatly benefit from human creativity. For example, evidence of trust derived from creative virtual evaluation scenarios can support the trust assurance of fast-paced runtime adaptation of intelligent behavior. Following this vision, in this paper, we introduce a methodological and architectural concept that interplays creative and social aspects of gaming into software engineering activities, more precisely into a virtual evaluation of system behavior. A particular trait of the introduced concept is that it reinforces cooperation between technological and social intelligence.**

***Index Terms*—Trust, Virtual Evaluation, Gaming**

I. Introduction

The systematic adoption of AI solutions is envisioned to uplift the human responsibilities in the emergent *feeling economy* [1] with an increased assignment of repetitive and analytical human cognitive processes to AI components, i.e. smart agents. Such a transition leaves human workers a free space for addressing more interpersonal, empathetic and ethical tasks.

However, from an engineering perspective there are ongoing challenges that hinder the trust into AI performance during operation. Particularly challenging from a trustworthiness perspective are intelligent systems that learn continuously during operation. While a different set of outputs provided for the same set of inputs may be an evidence of improved behaviour (due to continuous learning), it can also be a sign of sporadic malicious behavior that only activates in specific situations when a catastrophic impact is foreseeable. In our previous work [2] we have proposed a solution for ensuring a trusted execution of software smart agents based on runtime behavior prediction within a safe environment. The prediction is performed without executing the software that can potentially contain malicious behavior, but by executing its digital twins in a trusted real-time environment. In this paper we bring forward the idea of trust assurance of intelligent (and continuously evolving) behavior, by addressing the decision challenge of runtime behavioral control of a system. When a behavior or system component cannot be activated due to doubts regarding its trustworthiness, a decision needs to be taken whether or not to activate alternative trusted behaviors or components in the spirit of a fail operational scheme.

Although general to any system under the control of software smart agents, we'll exemplify our concept for the domain of safety critical systems, allowing us to be more specific in the explanations. For example, when an autonomous vehicle, which is a safety critical system, is faced with evidence of possible untrusted intentions of a software smart agent, it needs to decide very fast on a trusted course of action. If the vehicle driving around a school area relies on the intelligent activation of speed limits and the responsible behaviour is deceived by a malicious attack, the vehicle needs to detect this to trigger a fail-over behavior early enough in order to avoid accidents. Key to the envisioned scheme is the dynamic acquisition of trust evidence. Such evidence can be provided by humans that exercise system behaviors in creative settings within a design and runtime co-engineering framework. Based on evidence of challenging operational contexts collected from the field, creative explorations of behavioral variants of systems can be performed upfront. Such evidence, can support runtime activation of intelligent behavior. Crowd intelligence is in our opinion an immediate and scalable resource that enables a creative evaluation of intelligent behavior. To support this transition, in the current paper we propose an architectural and methodological concept that supports the outsourcing of behavioral evaluation in open scenarios to crowd intelligence through gaming.

In what follows, Section II presents an overview of the emerging transitions and trends on the roles of humans in the creative process of virtual evaluation. Section III introduces our methodological approach for using gaming as a resource to create trust during runtime with an architectural concept of platform presented in Section IV. Section V presents discussions and conclusions together with ongoing work and future research directions.

II. Emerging Transitions

A. Crowd Intelligence and AI Intelligence

Basing system intelligence on crowdsourcing, crowd intelligence [3] can through motivational schemes engage the large population into performing intelligent tasks, such as image recognition [4]. Engagement of population in supporting AI evolution is traditionally enforced through monetary rewards which build on extrinsic human motivation.

An emerging trend of gaining human engagement is further on brought through gamification schemes [5] which builds on the intrinsic motivation of a human being. Through an advanced level of commitment used in competition-based crowdsourcing platforms, unidirectional technological solutions are developed.

We believe that an uplift of the traditional crowd-sourcing concept can further on construct a sustainable AI-socio-technical evolution through integration of social and psychological human aspects shaped by social experts.

B. From machine to human readable specification scenarios

The advancements of autonomous vehicles rely on development of intelligent control trained on huge datasets. Large scale training is expected to improve reactions of autonomous systems to certain situations. But despite large data sets, accidents still occur when the training data does not cover all situations [6]. In order to fix this, development of corner case detection aim at identifying untypical situations [7], with an emerging trend of exploring human creativity in this direction. For example, frameworks such as [8] enable derivation of as many test scenarios as possible for autonomous driving, by closing the gap between machine-readable representations and human understanding.

On top of this, languages such as M-SDL (Measurable Scenario Description Language) [9] allow a simplified capture and reuse of scenarios, enabling specification of a mixture of conditions with the scope of identifying unknown hazards and edge cases for which an autonomous behavior can be safeguarded at runtime. Specifications that result from test scenarios then become requirements that guide the development of intelligent behavior.

Elevating from the idea of human understandable description of virtual evaluation scenarios, our approach also envisions the availability and readiness of solutions in a gaming setting at the convenience of the crowd.

C. Enriching the input domain

The dynamic acquisition of human-generated evidence of trust in open runtime environments can still be time consuming. Therefore, we propose enriching the variety of valid and invalid input through usage of techniques capable to manipulate available data sets in order to generate new inputs for valid solution. Usually applied for assessing the effectiveness of a testing approach, mutation analysis [10] is a commonly adopted approach for input transformations. In mutation testing, a mutant is a slightly modified version of the program under test, differing from it by a small, syntactic change. The underlying assumption of mutation testing and the coupling effect, is that, by looking for simple syntactic modifications, more complex, but real, faults can be found.

Creating new inputs by applying semantic information-preserving transformations is a challenge in different software research areas. As analyzed in [11] different approaches can be adopted such as: the metamorphic transformations focused on input alterations that mimic the environment conditions or real-world phenomena; the application of search-based approaches for eliciting collision scenarios; the exploitation of the boundaries of the input space so as to maximize the transitions in the behavior; or the investigation of the adversarial inputs able to trigger misbehaviors often very unlikely or impossible to occur in reality. In our concept we envision the adoption of mutation approaches to enrich the dynamic acquisition of trust evidence.

III. Gaming for Trust

During the design of autonomous processes, such as autonomous driving, different types of AI components are envisioned to either automate parts of the vehicle control or provide increased awareness of the runtime operational context. Typically, AI components are trained on data sets at design time. Then, during operation, the degree of matching between new situations and previously trained situations provides a level of trust into planned actions. Particularly challenging however are those situations, for which a trusted action that was initially decided needs to be modified due to an unforeseen event, e.g. the sudden detection of an obstacle. These situations stress the reactions speed of an autonomous vehicle and moreover has the potential to reduce the vehicles ability to keep an operational state. For example, a vehicle making a right turn at an intersection might need to react quickly to a sudden overtaking of another vehicle that is approaching from the opposite direction. Stopping the vehicle as a result of an immediate fail-over behavior will not only decrease the human trust into the autonomous driving capabilities but in this risky situation, it might not avoid an immediate crash. Instead, the vehicle should e.g. drive to the far right side of the street, which implies activation of another trusted behavior. Trust can be supported by evidence from previous similar evaluation scenarios in which the system did prove correct decisions and behavior.

A high degree of variations for the scenario configurations can be derived during design time by applying specific mutation operators to scenarios objects and data. Specifically, equivalence classes can be defined for context awareness or for related game objects that represent AI. These can be further used for deriving new equivalent scenarios for assessing the established level of trust and derivation of additional unexpected scenarios for behavioral evaluation. For example, considering a scenario in which a vehicle reacts quickly to sudden overtaking of a vehicle: i) different equivalent scenarios can be derived either by changing the type of weather condition or the type of road surface. In both cases the vehicle reaction should not be largely conditioned by the applied mutations, i.e., the vehicle should in any case drive to the far-right side of the street, but with a slight adaptation of speed. Alternatively, additional unexpected scenarios could be derived by mutating the overall situation. For instance substituting the *overtaking of a vehicle* with *approaching of blind curve* or ongoing earthquake. In both cases the vehicle reaction should not be the same as before. In this situations, the vehicle should not only keep the right side of the street but should also drastically reduce

the speed in order to avoid possible hazards. In both cases, previous evaluation of complex scenes can be opportunely mutated for providing evidences trusted behavior.

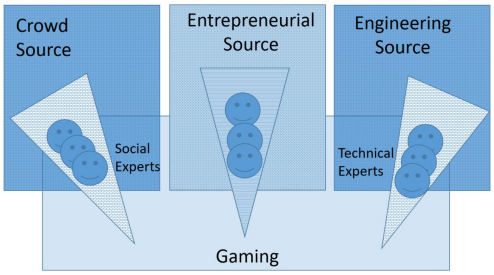

Fig. 1. High level view of the methodological approach

Fig. 1 presents a high level view of our proposed approach. Efficiency of behavior evaluation is achieved through creative gaming while its effectiveness is supported by a systematic mapping between the engineering world and the gaming world. The engineering world can be represented by real devices within a lab or virtual entities such as simulation models, and diagrams within a virtual testing environment. Psychological benefits for the population are assured through an interplay of social experts with the role of framing gaming scenes for the psychological needs of the crowd and at the same time leverage gamification schemes for engagement. We based this decision on recent developments within the gaming industry [12]. Even though until recently the scope of systematic triggering of emotions has been directed towards achieving business gains, recently, the gaming industry has been approved to produce games for the psychological benefits of humans.

Fig. 2 depicts the main components of a framework that enables the mapping between the engineering and gaming world. The virtual evaluation within an engineering setting needs to be framed into game scenes. Each scene contains multiple gaming objects which are mappings of systems, system components and technologies, such as AI components. This mapping is performed by engineers and technical experts, and can be blinded if the player is not aware of the representation of the game object in the real world, or unblinded if the game object represents concrete systems. Blinded mapping enables exploration of concepts in a creative manner and relies on logical mapping only, not structural mapping. For example, different types of wireless technologies can be mapped to different types of strings that connect two objects.

Unblinded mapping enables explorations with new concepts through an accurate representation of real world objects. Configurations from explorations within gaming scenes are then passed to a co-simulation framework which executes simulation models of systems and system components in various scenarios.

Evidence that supports trusted deviations of behavior within specific technical situations are then shipped on the real system as blueprints. The blueprints can describe trusted reconfiguration in a set of scenarios. The degree of trust results from

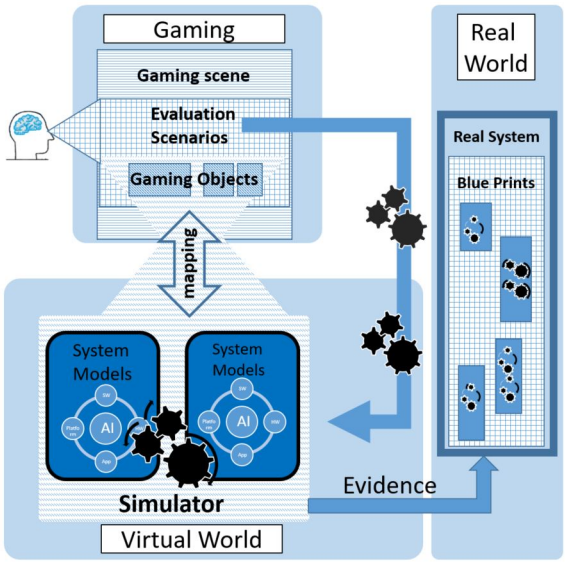

Fig. 2. Interplay of Virtual and Technical World

evidence within virtual evaluation scenarios and guide the real-time decision process. Based on sufficient evidence of scenario evaluation in the simulated environment, during runtime, the blueprints assign according levels of trust on the planned course of action.

IV. Prototype Platform

In this section, we exemplify the components of a platform that enables implementation of our methodological approach.

Fig. 3 outlines the architecture of such a platform in which *Game scenes* aggregate both *Data* and *Game Objects*. *Environmental Data* can be provided by simulators specialized in driving maneuvers and simulation of weather conditions, such as the LGSVL [13] which is Unity3D-based and Carla [14], an Unreal-based automotive simulator plugin. *Game objects* either represent only a logical mapping of engineering and technical concepts or accurately represent real-world objects. *Environmental Data* represents the inputs that a specific technology or system component is processing in a given scenarios, whereas *System Data* is provided by simulation models of *Virtual Platforms*. System data can consist of continuous or discrete values generated from the execution of simulation models. These values specify functional and non-functional properties of a system component, subject of virtual evaluation. The simulation models can be high level specifications models, detailed models defined in Simulink [15] or concrete implementations, including software implementation of *AI components*. In this way, within a game setting a deep learning algorithm, which is an AI component responsible for image recognition of an autonomous vehicle can be evaluated on a variety of input stimuli and provide a higher level of confidence in the intelligent reaction of the vehicle.

Each *Virtual Platform* is represented in the gaming world by a *Game Object* exercised within a *Game Scene*. The creative experimentation with game objects within a game

scene creates a variety of virtual *Evaluation Scenarios*. These scenarios are executed by a *Co-simulation framework* which integrates virtual platforms represented by simulation models. For enabling simulator interoperability, the integration of different simulation models within the co-simulation framework needs to be in conformance to standard interfaces, such as the FMI (Functional Mockup Interface) [16].

The evaluation scenarios within games specify *Configurations* that are further on exercised by the co-simulation framework. These configurations together with evidence of trust are integrated within *Blueprints* ready to be downloaded on systems.

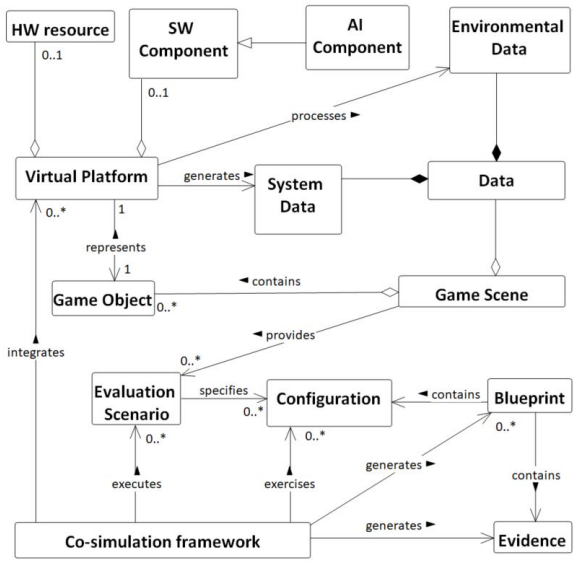

Fig. 3. Logical View of the Platform

V. Discussion and Conclusions

Elevating from the idea of human-understandable description of scenarios, the approach we have introduced, enables, through gaming availability and readiness of solutions a creative virtual evaluation of intelligent system behavior. Further on, with the support of a runtime simulation framework that we have introduced in [2], blueprints of trusted scenarios can increase the level of confidence into fast decisions with a direct support for agile adaptations to unforeseen runtime situations.

A. State of work and preliminary results

Our platform is based on FERAL simulator [17] and builds on the integration between virtual and real world, introduced in [18]. Additional research and engineering aspects that enable a runtime and design time co-engineering of trusted behavior have been introduced in [19].

B. Future and ongoing work

The scope of the entire concept encompasses many interesting research questions for further investigation, such as (a) definition of structural description of blue prints that support real-time decision control, (b) definition of blueprints for use cases within different domains, and (c) specification of evidence that supports the time criticality of runtime control activation.

Acknowledgment

This work is co-funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952702 (BIECO) 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] M.-H. Huang, R. Rust, and V. Maksimovic, "The feeling economy: managing in the next generation of artificial intelligence (ai)," *California Management Review*, vol. 61, no. 4, pp. 43–65, 2019.

[2] E. Cioroaica, T. Kuhn, and B. Buhnova, "(Do not) trust in ecosystems," in *Proceedings of the 41st International Conference on Software Engineering: New Ideas and Emerging Results*. IEEE Press, 2019, pp. 9–12.

[3] K. Xin, S. Zhang, X. Wu, and W. Cai, "Reciprocal crowdsourcing: Building cooperative game worlds on blockchain," in *2020 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2020, pp. 1–6.

[4] D. P. Sullivan, C. F. Winsnes, L. Åkesson, M. Hjelmare, M. Wiking, R. Schutten, L. Campbell, H. Leifsson, S. Rhodes, A. Nordgren *et al.*, "Deep learning is combined with massive-scale citizen science to improve large-scale image classification," *Nature biotechnology*, vol. 36, no. 9, pp. 820–828, 2018.

[5] C. Yang, H. J. Ye, and Y. Feng, "Using gamification elements for competitive crowdsourcing: exploring the underlying mechanism," *Behaviour & Information Technology*, pp. 1–18, 2020.

[6] Y. Tian, K. Pei, S. Jana, and B. Ray, "Deeptest: Automated testing of deep-neural-network-driven autonomous cars," in *Proceedings of the 40th international conference on software engineering*, 2018, pp. 303–314.

[7] J. Bolte, A. Bar, D. Lipinski, and T. Fingscheidt, "Towards corner case detection for autonomous driving," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 438–445.

[8] "Foretellix," https://www.foretellix.com/open-language/, [Online; accessed 12-December-2020].

[9] "Foretellix M-SDL," https://www.foretellix.com/category/m-sdl/, [Online; accessed 12-December-2020].

[10] Y. Jia and M. Harman, "An analysis and survey of the development of mutation testing," *IEEE Transactions on Software Engineering*, vol. 37, no. 5, pp. 649–678, Sept 2011.

[11] V. Riccio, G. Jahangirova, A. Stocco, N. Humbatova, M. Weiss, and P. Tonella, "Testing machine learning based systems: a systematic mapping," *Empirical Software Engineering*, vol. 25, no. 6, pp. 5193–5254, 2020.

[12] M. Anderson, "Prescription-strength gaming: Adhd treatment now comes in the form of a first-person racing game-[news]," *IEEE Spectrum*, vol. 57, no. 8, pp. 9–10, 2020.

[13] "LGSVL simulator," https://www.lgsvlsimulator.com/, [Online; accessed 12-December-2020].

[14] "Carla Simulator," https://carla.org/, [Online; accessed 04-December-2020].

[15] "Simulink," https://www.mathworks.com/products/simulink.html, [Online; accessed 02-December-2020].

[16] "FMI," https://fmi-standard.org/, [Online; accessed 06-December-2020].

[17] T. Kuhn, T. Forster, T. Braun, and R. Gotzhein, "Feral—framework for simulator coupling on requirements and architecture level," in *Formal Methods and Models for Codesign (MEMOCODE), 2013 Eleventh IEEE/ACM International Conference on*. IEEE, 2013, pp. 11–22.

[18] E. Cioroaica, T. Kuhn, and T. Bauer, "Prototyping automotive smart ecosystems," in *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*. IEEE, 2018.

[19] E. Cioroaica, S. Chren, A. Larsson, R. Chillarege, T. Kuhn, D. Schneider, C. Wolschke *et al.*, "Towards creation of automated prediction systems for trust and dependability evaluation," in *2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. IEEE, 2020, pp. 1–6.