Towards Explainable, Compliant and Adaptive Human-Automation Interaction

Barbara Gallina[1], Görkem Pacaci[2], David Johnson[2], Steve McKeever[2], Andreas Hamfelt[2], Stefania Costantini[3], Pierangelo Dell'Acqua[4] and Gloria-Cerasela Crisan[5]

[1] Mälardalen University, [2] Uppsala University, [3] University of L'Aquila
[4] Linköping University, [5]Vasile Alecsandri University of Bacau
barbara.gallina@mdh.se

Abstract. AI-based systems use trained machine learning models to make important decisions in critical contexts. The EU guidelines for trustworthy AI emphasise the respect for human autonomy, prevention of harm, fairness, and explicability. Many successful machine learning methods, however, deliver opaque models where the reasons for decisions remain unclear to the end user. Hence, accountability and trust are difficult to ascertain. In this position paper, we focus on AI systems that are expected to interact with humans and we propose our visionary architecture, called ECA-HAI (Explainable, Compliant and Adaptive Human-Automation Interaction)-RefArch. ECA-HAI-RefArch allows for building intelligent systems where humans and AIs form teams, able to learn from data but also to learn from each other by playing "serious games", for a continuous improvement of the overall system. Finally, conclusions are drawn.

Keywords: Programme Synthesis, Explainable AI, Compliant AI, Serious Games.

1 Introduction

Artificial Intelligence (AI) is increasingly being used in applications that impact society. To make important predictions/decisions in critical contexts, AI-based systems make use of trained machine learning (ML) models, which may consist of (deep) neural networks ((D)NN). For instance, an AI-based semi-autonomous-driving vehicle is expected to evaluate and "co-manage" the risk together with the driver, while a fully autonomous vehicle is even expected to self-manage the risk.

The demand for trustworthiness is increasing from the various stakeholders of AI. According to the guidelines proposed by the European Commission, trustworthy AI means guaranteeing compliance, safety, security, reliability, adaptability, explainability. This last guarantee, which sometimes is referred to as eXplainable AI (XAI), has been identified as an utmost need for the adoption of ML methods in critical contexts. The initially proposed monolithic end-to-end NN-based paradigm for self-driving vehicles, known as ALVINN [1], suffers from opacity. Subsequent proposals, introduced modularity and limited the role of ML. The current state of the art envisioned paradigm for self-driving vehicles, however, re-introduces an end-to-end NN-based solution [2], that is now modular and expected to enable supervision so as to become explainable. In general, beyond the automotive application of AI-based systems, many successful machine learning methods, however, still deliver opaque models where the reasons for decisions may remain inexplicable to the end user. This lack of transparency ensures that responsibility for decision making cannot be corroborated. Either one limits the computational reach of an AI artefact, or one significantly restricts the data on which the artefact is built to ensure compliance, or one tries to understand it.

In this position paper, we propose a novel architecture, called ECA-HAI-RefArch, for building intelligent systems where humans and AIs form teams, able to learn from data but also to learn from each other by playing "serious games", for a continuous improvement of the overall system. ECA-HAI-RefArch integrates and extends solutions for: explaining AI-based information systems, checking/arguing and self-reflecting about compliance of the explained AI behaviour with the normative spaces of pertinence as well as about the compliance of the interaction with the upcoming normative spaces, gamifying the interaction between the intelligent artificial system and the human intelligence.

The rest of this paper is structured as follows. In Section 2, we provide essential background information. In Section 3, we describe our architecture. In Section 4, we discuss related work. Finally, in Section 5, we draw our conclusions.

2 Background

In this section, we recall the background on which we build our proposed architecture.

2.1 Rule Induction of CNP Explanations (RICE)

The RICE method [3] generates explainable models through a combination of sensitivity analysis to extract input-output pairs that are critical to interpreting the black box's behaviour, followed by a program synthesis stage to generate an alternative representation of how the black box functions. Unlike other established explanation methods (such as LIME [4]), which provide localized explanations, RICE provides a globally interpretable explanation. RICE has three phases: 1) the *probing* phase takes the opaque model, the types of the inputs and outputs of the model, and generates a dataset of critical example input/output pairs; 2) the *synthesis* phase deals with searching the space of programmes to derive a mapping, namely a programme written in CNP (COMBILOG with Named Projection[4]), from the critical inputs to outputs. Finally, 3) the *interpretation* phase ensures that the CNP programme can be translated into human language or to other logic representations.

2.2 ACCEPT

ACCEPT (Automated Compliance Checking of Engineering Process plans against sTandards) [5-6] is a tool-supported method for modelling processes checkable for compliance, i.e., processes elements enriched with compliance information through annotations representing formalized standards requirements in FCL (Formal Contract Logic) [7]. FCL permits users to represent and reason about normative knowledge, i.e., the obligations and permissions can be defined and the compliance effects they produce in the process plans can be formally verified.

2.3 MDSafeCer

MDSafeCer (Model Driven Safety Certification) [8] is a model-driven tool-supported method for semi-automatically generating process-based arguments from fallacy-free process models. MDSafeCer generates structured arguments that link the evidence with the claims about compliance with the normative space. The arguments are generated once the absence of *omission of key evidence* is verified.

2.4 Reflection

Quoting Torresen et al [10]: "Self-aware and self-expressive computing describes an emerging paradigm for systems and applications that proactively gather information; maintain knowledge about their own internal states and environments; and then use this knowledge to reason about behaviours, revise self-imposed goals, and self-adapt. Systems that gather unpredictable input data while responding and self-adapting in uncertain environments are transforming our relationship with and use of computers." This kind of advanced self-aware reflective systems can be realized by means of the notion of Reflection Principle [11], by which a designer can encode various forms of uncertain or plausible reasoning, and sophisticated meta-constraints (either local or global) over the system's functioning [12-13] aimed at run-time self-checking.

2.5 Gamification

Gamification is *the use of game design elements in non-game contexts* [14]. It offers new approaches to adult learning, as it uses intrinsic motivation for achieving individual, team or social objectives [15]. Gamification could also be used for controlling artificial hybrid systems, where computational intelligence is improved by complementing it with human intelligence in an interactive ML approach [16].

3 A Vision towards ECA-HAI

In this section, we present our visionary architecture, called ECA-HAI RefArch, which stands for Explainable, Compliant and Adaptive Human-Automation Interaction Reference Architecture. ECA-HAI RefArch builds on top of the building blocks, which were introduced in Section 2. More specifically, as depicted in Fig. 1, the ECA-HAI RefArch consists of a two-layered architecture.

The first layer comprises the components used at design time: 1) a component that perfors the synthesis of an opaque neural network model (based on the RICE method); 2) a component that performs the interpretation of the Explained Neural Network Model (based on the RICE method in conjunction with ACCEPT and MDSafeCer) and presents the interpretation in terms of compliance results (the process-based behavioural representation of the neural network model complies with e.g. the *motor* *vehicle safety act* i.e., the neural network model must not lead to unreasonable risk of death or injury; the neural network model must not lead to responsibility delegation when inappropriate e.g. by delegating the responsibility to humans when humans cannot control an *hazardous event*) and argumentation fragments (fragments of justifications for the neural network model behaviour in relation to the stated goals). This interpretation ensures that the opaque artefact is both legally and ethically sound. The second layer comprises the components used at run-time: 1) a component that transforms the model of neural network into a representation adequate for serious games within a gamification environment (a virtual arena, where the interaction between the human and the artificial intelligence can safely take place and be explored without repercussions, by playing with 'what if' scenarios, by guiding the two learners and by tracking their performance); 2) a component that is responsible for a twofold functionality: the gamification of the interaction between the human (e.g., urban air traffic controller, road vehicle driver, etc.) and the artificial intelligence (represented by the neural network model) and the generation of a model describing the interaction and result of the learning experience during the serious game; 3) a component that (based on reflection/argumentation/quality evaluation) interprets the generated output regarding the interaction and produces: a compliance report, an argument for the assurance case to assure society regarding the harmless interaction, and a quality report regarding the learning experience. The dynamics of the architecture is given in terms of an activity-diagram-like style where the components are the activities.

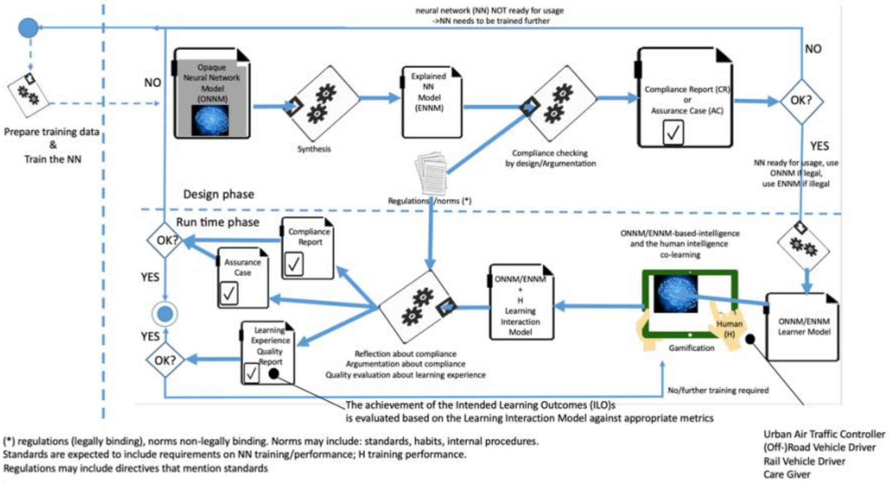
neural network (NN) NOT ready for usage
->NN needs to be trained further
NO
Opaque Neural Network Model (ONNM)
Synthesis
Explained NN Model (ENNM)
Compliance checking by design/Argumentation
Compliance Report (CR) or Assurance Case (AC)
✓
OK?
NO
YES
NN ready for usage, use ONNM if legal, use ENNM if illegal
Prepare training data & Train the NN
Regulations /norms (*)
Design phase
Run time phase
OK?
YES
Compliance Report
✓
Assurance Case
✓
YES
OK?
Learning Experience Quality Report
✓
Reflection about compliance
Argumentation about compliance
Quality evaluation about learning experience
ONNM/ENNM + H Learning Interaction Model
ONNM/ENNM-based-intelligence and the human intelligence co-learning
Human (H)
Gamification
ONNM/ENNM Learner Model
No/further training required
The achievement of the Intended Learning Outcomes (ILO)s
is evaluated based on the Learning Interaction Model against appropriate metrics
(*) regulations (legally binding), norms non-legally binding. Norms may include: standards, habits, internal procedures.
Standards are expected to include requirements on NN training/performance; H training performance.
Regulations may include directives that mention standards
Urban Air Traffic Controller
(Off-)Road Vehicle Driver
Rail Vehicle Driver
Care Giver

Fig. 1. ECA-HAI RefArch

4 Related work

To the best of our knowledge, no one has adequately studied human-automation interaction trust or its potential to be increased by means of such a progressive combination of approaches. Thus, we can claim that our proposed "serious games" go beyond XAI and pioneer X/C H-AI I (eXplainable/Compliant Human-Artificial Intelligence Interaction). With regards to ML trustworthiness and explainability, in [17] the authors provide a comprehensive survey on the opportunities and challenges of explanation extraction. Definitions of trustworthy AI, as elaborated in [18], rely on explainability of the artefacts in use. There are mainly two approaches to explainability. First is revealing the specific sections of a case that lead to a decision, as LIME does. Second is the set of methods that produce a general description of the opaque AI artefact [19]. With regards to argumentation about AI-based systems, proposals for arguing about trustworthy ML have been proposed by various authors [20-21]. However, these proposals lack a holistic perspective, limiting the focus on specific domains, or concern. With regards to compliance checking in the context of AI-based systems, in [22] authors discuss how the combination of mental attitudes and obligations can be framed in Defeasible Logic and how this logic permits users to reason about norm-compliant artificial intelligence. With regards to "serious games", in [23], authors propose a game theoretic traffic model that can be used to test and compare various autonomous vehicle decision and control systems and calibrate the parameters of an existing control system.

5 Conclusion and future work

In this paper, we have presented our vision for building intelligent systems where humans and AIs form teams, able to learn from data but also to learn from each other by playing "serious games", for a continuous improvement of the overall system so that subsequent refinements will yield a responsible AI. The dichotomy of mind/AI is speculated in a way that provides reciprocal skill development and better understanding of each other's' role and performance. Continuous training using different human experts offers to our envisioned solution a potentially continuous upgrade and adaptation to new events and scenarios (including edge cases). By exposing AI to more and more human intelligence, the hybrid team will become more and more effective (rationality-&-creativity-based synergies could emerge and could be detected and used to develop future normative spaces). As Marvin Minsky stated: "What magical trick makes us intelligent? The trick is that there is no trick. The power of intelligence stems from our vast diversity, not from any single, perfect principle." [24]. As future work, we intend to make our vision concrete.

References

[1] Pomerleau D. A., ALVINN: An Autonomous Land Vehicle in a Neural Network. Advances in neural information processing systems. Morgan Kaufmann, pp. 305-313, (1989).

[2] Luo, W., Yang, B., and Urtasun, R. Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting with a Single Convolutional Net. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, pp. 3569-3577, (2018).

[3] Paçacı, G., Johnson, D., McKeever, S., Hamfelt, A.: "Why Did You Do That?" Explaining black box models with Inductive Synthesis. Computational Science-ICCS. LNCS, vol. 11536, pp. 334-345 (2019).

[4] Paçacı, G., McKeever, S., Hamfelt, A.: Compositional Relational Programming with Name Projection and Compositional Synthesis. Perspectives of System Informatics. LNCS, vol. 11964, pp. 306-321 (2017).

[5] Castellanos Ardila, J. P., Gallina, B.: Separation of Concerns in Process Compliance Checking: Divide-and-Conquer. 27th European & Asian System, Software & Service Process Improvement & Innovation. pp. 1-12 (2020)

[6] Castellanos Ardila, J. P., Gallina, B., Ul Muram, F.: Enabling Compliance Checking against Safety Standards from SPEM 2.0 Process Models. Euromicro Conference on Software Engineering and Advanced Applications. pp. 45 - 49 (2018).

[7] Governatori, G.: Representing Business Contracts in RuleML. International Journal of Cooperative Information Systems. pp. 181-216 (2005)

[8] Gallina, B. A Model-driven Safety Certification Method for Process Compliance. 2nd International Workshop on Assurance Cases for Software-intensive Systems (ASSURE), Naples, Italy. IEEE, pp. 204-209, (2014).

[9] Ul Muram, F., Gallina, B., Gomez Rodriguez, L.: Preventing Omission of Key Evidence Fallacy in Process-based Argumentations. 11th International Conference on the Quality of Information and Communications Technology (QUATIC), IEEE, Coimbra, Portugal, September, (2018).

[10] Torresen, J., Plessl, C. and Yao, X. Self-Aware and Self-Expressive Systems. *Computer*, vol. 48, no. 07, pp. 18-20, (2015).

[11] Barklund, J., Costantini, S., Dell'Acqua, P., Lanzarone, G.A.: Reflection Principles in Computational Logic. Journal of Logic and Computation, vol. 10, pp 743-786, (2000).

[12] Costantini, S., and Formisano, A.: Adding Metalogic Features to Knowledge Representation Languages. Fundamenta Informaticae, to appear, (2020).

[13] Costantini, S.: Ensuring Trustworthy and Ethical Behavior in Intelligent Logical Agents. 35th Italian Conference on Computational Logic, to appear (2020).

[14] Deterding, S., Dixon, D., Khaled, R., & Nacke, L.: From game design elements to gamefulness: Defining “gamification”. 15th International Academic MindTrek Conference: Envisioning Future Media Environments, pp. 9–15, New York, NY, USA, ACM, (2011).

[15] Mitchell, R; Schuster L; Jin H.S.: Gamification and the impact of extrinsic motivation on needs satisfaction: Making work fun? Journal of Business Research, vol. 106, pp. 323-330, (2020).

[16] Holzinger, A., Plass, M., Kickmeier-Rust, M., Holzinger, K., Crişan, G.C., Pintea, C.M.; Palade, V.: Interactive machine learning: experimental evidence for the human in the algorithmic loop. Applied Intelligence, vol. 49, pp. 2401-2414, (2019).

[17] Barredo-Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, vol. 58, pp.82-115, (2020).

[18] Ribeiro, M. T., Singh, S., Guestrin, C.: “Why should I trust you?” Explaining the predictions of any classifier. 22nd ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 1135-1144, (2016).

[19] Biran, O., Cotton, C.: Explanation and justification in machine learning: A survey. IJCAI-17 Workshop on eXplainable AI (XAI), vol. 8, no. 1, pp. 8-13, (2017).

[20] Burton, S., Gauerhof, L., Heinzemann, C.: Making the case for safety of machine learning in highly automated driving. Computer Safety, Reliability, and Security, LNCS, vol. 10489, pp. 5–16, (2017).

[21] Bloomfield, R., Khlaaf, H., Ryan Conmy, P., Fletcher, G.: Disruptive Innovations and Disruptive Assurance: Assuring Machine Learning and Autonomy. Computer, vol. 52, no. 9, pp. 82-89, (2019).

[22] Governatori, G., Rotolo, A. BIO logical agents: Norms, beliefs, intentions in defeasible logic. Autonomous Agents Multi-Agent Systems, vol. 17, no. 1, pp. 36–69, August, (2008).

[23] Li, N., Oyler, D. W., Zhang, M., Yildiz, Y., Kolmanovsky, I., and Girard, A. R.: Game Theoretic Modeling of Driver and Vehicle Interactions for Verification and Validation of Autonomous Vehicle Control Systems. IEEE Transactions on Control Systems Technology, vol. 26, no. 5, pp. 1782-1797, September (2018).

[24] Minsky, M.L.:The society of mind, Ed. Simon and Schuster, ISBN 0-671-60740-5, (1986).