Towards Explainable Security for ECA Rules

Bernardo Breve[1,*,†], Gaetano Cimino[1,*,†] and Vincenzo Deufemia[1,†]

[1]University of Salerno, via Giovanni Paolo II, Fisciano (SA), 84084, Italy

Abstract

With the rise in popularity of smart objects and online services, the use of Trigger-Action Platforms for the definition of custom behaviors is growing significantly. These platforms enable end-users to create Event-Condition-Action (ECA) rules for triggering actions upon event occurrences on physical devices or online services in different domains. ECA rules could easily expose end-users to security risks mainly due to their low level of knowledge and awareness. To alleviate this problem, classification models can be used for identifying possible security issues that ECA rules could inflict when triggered. However, the results produced by these classifiers may not be understood by end-users. This position paper provides first insights concerning the application of AI models for generating natural language explanations according to the identified risks of ECA rules.

Keywords

trigger-action platforms, ECA rules, security, explainable AI

1. Introduction

The Internet of Things (IoT) is an innovative paradigm that allows end-users to put smart objects and online services in direct communication with each other. Recently, we have witnessed the rise of many platforms that facilitate the interaction between end-users and their ecosystems. Among them, the most popular ones follow the *trigger-action programming* (TAP) paradigm, which empowers end-users to create automation through conditional rules. In particular, these rules are based on the *Event-Condition-Action* (ECA) paradigm, where an action is performed when an event that satisfies a specific condition is triggered. From this perspective, it is easy to understand why such rules are massively employed. In fact, non-expert end-users can avoid worrying about low level technical details and focus their attention on the custom behaviors they want to define. However, it is necessary to consider that most end-users using these platforms do not have a technical or security background. This lack of knowledge could lead end-users to define rules that might compromise their privacy or the security of their environment. For instance, at first sight, the rule "*Keep your Facebook and Twitter profile pictures in sync*" could not seem harmful. However, an end-user may forget that such a rule is active, allowing the upload of an unwanted photo to Twitter, causing possible embarrassment. Thus, it is essential to identify these situations in order to warn end-users about the risks they expose themselves.

EMPATHY: 3rd International Workshop on Empowering People in Dealing with Internet of Things Ecosystems. Workshop co-located with AVI 2022, June 06, 2022, Frascati, Rome, Italy.

*Corresponding author.

†These authors contributed equally.

bbreve@unisa.it (B. Breve); gcimino@unisa.it (G. Cimino); deufemia@unisa.it (V. Deufemia)

0000-0002-3898-7512 (B. Breve); 0000-0001-8061-7104 (G. Cimino); 0000-0002-6711-3590 (V. Deufemia)

Many works have addressed these issues by carefully categorizing the types of harm that could be inflicted [1, 2, 3, 4]. An approach for their detection is the use of Artificial Intelligence (AI) techniques, since they allow for analyzing the semantic and contextual information in which an ECA rule is applied. In [5], we firstly evaluated the feasibility of classifying ECA rules by training a classification model on manually labeled rules with respect to four classes of risk [2]. Such a solution provides many advantages over static approaches, such as the analysis of the information flow [2], which performs a static services analysis without considering the context where a rule is employed.

The results produced by classification models are difficult to understand without expert knowledge. Thus, it is fundamental to provide end-users with valid explanations describing the risks connected to a rule in a comprehensible manner, with the aim of enhancing the end-users trust. Such a task can be faced by employing the *Explainable Security* (XSec) paradigm, which is inspired by the *eXplainable Artificial Intelligence* (XAI) research field [6].

This position paper highlights existing AI techniques that can be used for generating natural language explanations clarifying why an ECA rule might cause harm.

2. Generation of explanations

Over the last few years, the need for solutions supporting end-users in understanding the results of AI techniques is progressively increasing. To this end, the Defense Advanced Research Projects Agency proposed the XAI paradigm [6], which deals with elucidating, in total transparency, the reason behind the outputs of an AI model.

Many XAI-based solutions have been proposed in the literature. For instance, Ribeiro *et al.* introduced Local Interpretable Model-Agnostic Explanations (LIME) [7], a technique that can faithfully explain the results of any AI model using a linear model. More specifically, we can use LIME to plot information about a model prediction, such as the probability distribution over target classes, the relevance of each feature of an instance in the classification task, and the set of words in a sentence that leads the model towards a specific decision. Similarly, Lundberg *et al.* presented SHapley Additive exPlanations (SHAP) [8], an additive feature attribution method based on Shapley values and game theory. SHAP makes it possible to identify the predictive power of features and the relationships between them and the target class.

Although these solutions allow for increasing end-user confidence and trust in prediction models by providing knowledge that can be easily interpreted, they may be ineffective for the audience we consider. Indeed, it is worth noting that such solutions arise intending to define visual representations that facilitate end-users to comprehend model decisions in producing a specific output (e.g., which features lead to a prediction rather than another). Instead, in the context of ECA rules, we need to explain to end-users the scenarios in which a rule could become dangerous for their privacy or security.

Inspired by the XAI paradigm, Viganò *et al.* proposed a new paradigm: XSec [9]. The latter aims to explain reasoning about privacy and security vulnerabilities and concrete attacks on systems. The authors suggest several approaches to produce different explanations according to the levels of detail required by the stakeholders. For instance, explanations for a system analyst might be presented with a low level of abstraction, such as explanation trees, attack

trees, formal languages, and so forth. Instead, non-expert end-users need easily understandable explanations, so a higher level of abstraction must be adopted.

A possible solution is using natural language explanations, which are immediately understandable by all types of end-users and can be customized according to the context of the use of rules. Figure 1 shows an example of the process yielding the generation of explanations for risky ECA rules. In particular, features concerning an ECA rule, i.e., event, condition, and action, are considered by a classification module, which identifies the class of risk. Finally, the XAI module will generate an explanation by considering both the ECA rule characteristics and the identified risk.

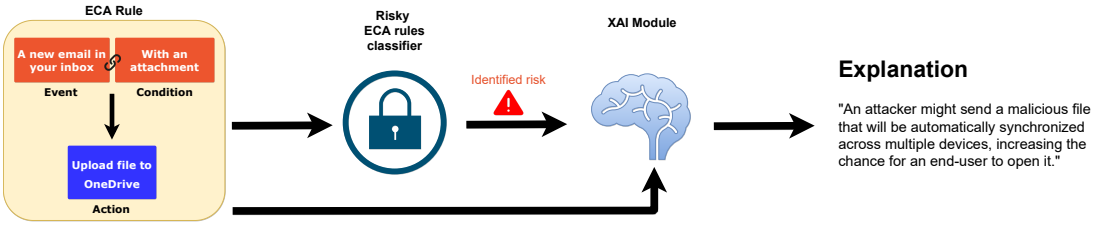

Figure 1: An example of the flow for generating explanations concerning risky ECA rules

The automatic generation of textual explanations is faced by another branch of XAI, called Explainable AI with *Natural Language Explanations* (Natural-XAI) [10]. The latter aims to build AI models capable of generating sentences that justify the ground-truth predictions inferred by classification models. Several approaches could be adopted to generate textual explanations that express the reason involving the risks associated with ECA rules. Among them, language models represent an effective way to achieve this goal since they can produce a structured text relating the risk and the rule's context of use. The existing models differ in the way they manipulate the data. For instance, Text-to-Text-Transfer-Transformer (T5) [11] is a model that can generate a sentence by taking a list of keywords as input. Thus, it is required to employ a further module for extracting the keywords from the rules' information. On the other hand, a model as Generative Pre-trained Transformer 2 (GPT-2) [12] requires the definition of a specific format for the rules to produce the textual explanations corresponding to the risks. Concerning this matter, one approach that could be exploited is *prompt-based learning* [13]. The latter focuses on finding the most appropriate prompt to adopt with a language model in order to manipulate its behavior and predict the desired output. Recently, models capable of jointly generating the prediction and explanation for a given instance are becoming increasingly popular [14, 15]. These models provide an advantage with respect to the necessity of applying in combination two different models, i.e., a classifier and a language model.

At the workshop, we will present how language models can be adopted to achieve the discussed goals.

Acknowledgments

This work has been supported by the Italian Ministry of University and Research (MUR) under grant PRIN 2017 "EMPATHY: Empowering People in deAling with internet of THings ecosYstems"

(Progetti di Rilevante Interesse Nazionale — Bando 2017, Grant 2017MX9T7H).

References

[1] I. Bastys, M. Balliu, A. Sabelfeld, If this then what? Controlling flows in IoT apps, in: Proceedings of ACM SIGSAC Conference on Computer and Communications Security, ACM, 2018, p. 1102–1119.

[2] M. Surbatovich, J. Aljuraidan, L. Bauer, A. Das, L. Jia, Some recipes can do more than spoil your appetite: Analyzing the security and privacy risks of IFTTT recipes, in: Proceedings of the 26th International Conference on World Wide Web, 2017, p. 1501–1510.

[3] C. Cobb, M. Surbatovich, A. Kawakami, M. Sharif, L. Bauer, A. Das, L. Jia, How risky are real users' IFTTT applets?, in: Proceedings of the Sixteenth USENIX Conference on Usable Privacy and Security, USENIX Association, USA, 2020, pp. 505–529.

[4] Q. Wang, P. Datta, W. Yang, S. Liu, A. Bates, C. A. Gunter, Charting the attack surface of trigger-action iot platforms, in: Proceedings of the 2019 ACM SIGSAC conference on computer and communications security, 2019, pp. 1439–1453.

[5] B. Breve, G. Cimino, V. Deufemia, Towards a classification model for identifying risky IFTTT applets, in: G. Desolda, V. Deufemia, M. Matera, F. Paternò, M. Zancanaro, F. Vernero (Eds.), Proceedings of the 2nd International Workshop on Empowering People in Dealing with Internet of Things Ecosystems co-located with with INTERACT 2021, Bari, Italy, Online / Bari, Italy, September 30, 2021, volume 3053 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 33–37.

[6] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, Xai—explainable artificial intelligence, Science Robotics 4 (2019) eaay7120.

[7] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[8] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[9] L. Vigano, D. Magazzeni, Explainable security, in: 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE, 2020, pp. 293–300.

[10] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom, e-snli: Natural language inference with natural language explanations, Advances in Neural Information Processing Systems 31 (2018).

[11] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, arXiv preprint arXiv:1910.10683 (2019).

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[13] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, arXiv preprint arXiv:2107.13586 (2021).

[14] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, T. Darrell, Generating visual explanations, in: European conference on computer vision, Springer, 2016, pp. 3–19.
[15] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, M. Rohrbach, Multimodal explanations: Justifying decisions and pointing to the evidence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8779–8788.