Explainable Security (position paper)*

Luca Viganò and Daniele Magazzeni

Department of Informatics
King's College London, London, UK
luca.vigano@kcl.ac.uk, daniele.magazzeni@kcl.ac.uk

The security of information, data, processes, software, protocols, computers, networks and systems is notoriously a challenging problem (and very often an undecidable one). Security is difficult. It is difficult to achieve, to reason about, to apply, to understand, to teach. It is difficult to *explain*.

The Defense Advanced Research Projects Agency (DARPA) recently launched the *Explainable Artificial Intelligence (XAI)* program that aims to create a suite of new AI techniques that enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems. Some research on explainable AI had already been published before DARPA's program (e.g., [13,11,6]), but XAI encouraged a large number of researchers to take up this challenge. In the last couple of years, several publications have appeared that investigate how to explain the different areas of AI, such as *machine learning* [7], *recommender systems* [9], *robotics* and *autonomous systems* [12], *constraint reasoning* [5] and *planning* [3,4].[1]

Inspired by the XAI program, we propose a new paradigm in security research:

Explainable Security (XSec).

Some pioneering works on explaining security have focused on *security for relational databases* [2] and on *explanation and trust* [10]. In [2], Bender, Kot and Gehrke propose a new model in which policy decisions are explainable. In this model, instead of simply rejecting an unauthorized query by a principal, the system provides the principal with a concise explanation of why the query was rejected and what additional permissions the principal would need to be granted for a successful execution. The principal can then refine the query or request additional permissions based on the explanation provided. In [10], Pieters investigates the relation between explanation and trust, focusing in particular on expert systems and e-voting systems. He discusses two main goals that an explanation may have: *transparency* (e.g., to allow users to understand what the designers have done to protect them) and *justification* (e.g., offering reasons for an action). He contrasts *explanation-for-trust* (i.e., explanation of how a system works, by revealing details of its *internal* operations) with *explanation-for-confidence* (i.e., explanation to make the user feel comfortable in using the system, by providing information on its *external* communications).

We argue that XSec is a difficult problem and that it has unique and complex characteristics. In fact, XSec is more complex than what is discussed by Bender et al. and by Pieters; it is also tightly connected to, but different from, *usable security* [15], *security awareness* [16] and *security economics* [1]. This is because XSec involves several different stakeholders (i.e., the system's developers, analysts, users and attackers) and is multi-faceted by nature (as it requires reasoning about system model, threat model and properties of security, privacy and trust as well as about concrete attacks, vulnerabilities and countermeasures).[2]

* A preliminary, but longer, version of this paper was presented as a poster (without publication) at the "IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence (XAI)" [14]. We thank David W. Aha, Fabio Mercorio, Diego Sempreboni and the anonymous reviewers of the XAI 2018 workshop for their useful comments and suggestions.

[1] See also [8] and `http://home.earthlink.net/~dwaha/research/meetings/faim18-xai/`.

[2] One could argue, e.g., that in order to be usable security needs first to be explainable, or vice versa. However, there is no shortage of real-life security systems and solutions that are usable without being explainable or without explaining themselves. Think about the most recent smartphones, which come without any user manual. They are usable but provide little explanation so that often users don't understand why and how they work, with possible dire consequences for the ultimate security of the data that the smartphone will manage. Similarly, explanations will need to be understandable by the human users to contribute to usability.

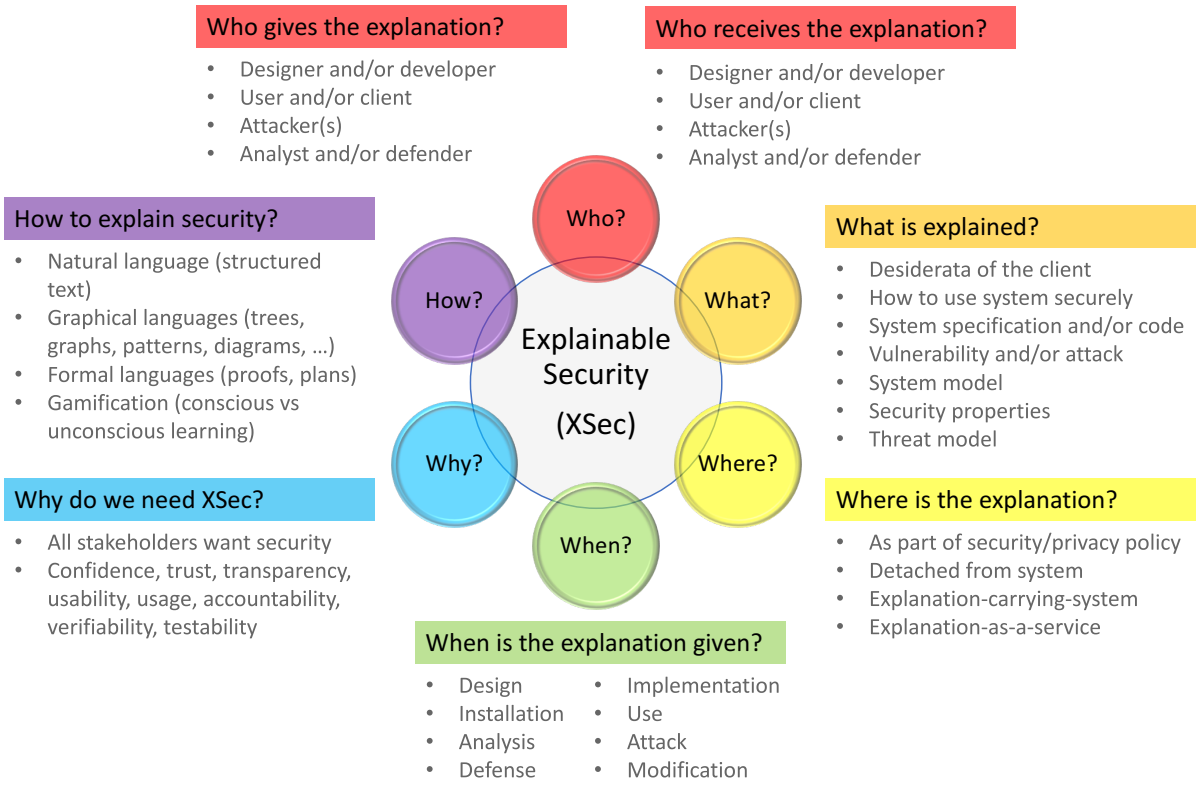

Fig. 1: The Six Ws of Explainable Security

Figure 1 shows the "Six Ws" of XSec (Who? What? Where? When? Why? and How?). We propose to use this figure as a roadmap to identify possible research directions, and to describe the challenges of XSec and how they could be tackled.

We have already begun a more detailed investigation of the different techniques and tools that can be exploited or need to be developed to answer the questions posed by the different Ws, as well as formalizing the relationships and interdependences between the Ws. To that end, we plan to capitalize on the growing amount of results on XAI together with the pioneering research on explanations in security and trust. We also plan to explore in more detail the connections and synergies between XSec and formal methods, argumentation and planning for security, as well as with usable security, security awareness and security economics.

References

1. Anderson, R.: Economics and Security Resource Page (2018), `http://www.cl.cam.ac.uk/~rja14/econsec.html`
2. Bender, G., Kot, Ł., Gehrke, J.: Explainable Security for Relational Databases. SIGMOD (2014)
3. Chakraborti, T., Sreedharan, S., Zhang, Y., Kambhampati, S.: Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. IJCAI (2017)
4. Fox, M., Long, D., Magazzeni, D.: Explainable Planning. IJCAI Workshop on Explainable Planning (2017)
5. Freuder, E.: Explaining ourselves: Human-aware constraint reasoning. AAAI (2017)
6. Gedikli, F., Jannach, D., Ge, M.: How Should I Explain? A Comparison of Different Explanation Types for Recommender Systems. International Journal of Human–Computer Studies 72, 367–382 (2014)
7. Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., Darrell, T.: Generating visual explanations. ECCV (2016)
8. Miller, T.: Explanation in Artificial Intelligence: Insights from the Social Sciences. Artif. Intell. 267 (2019)

9. Muhammad, K., Lawlor, A., Smyth, B.: Explanation-based Ranking in Opinionated Recommender Systems. AICS (2016)
10. Pieters, W.: Explanation and trust: What to tell the user in security and AI? Ethics Inf Technol 13 (2011).
11. Seegebarth, B., Müller, F., Schattenberg, B., Biundo, S.: Making Hybrid Plans More Clear to Human Users - A Formal Approach for Generating Sound Explanations. ICAPS (2012)
12. Sheh, R.: "Why did you do that?" Explainable intelligent robots. AAAI Workshop on Human-Aware Artificial Intelligence (2017)
13. Sohrabi, S., Baier, J.A., McIlraith, S.A.: Preferred Explanations: Theory and Generation via Planning. AAAI (2011)
14. Viganò, L., Magazzeni, D.: Explainable Security. http://arxiv.org/abs/1807.04178 (2018),
15. Wash, R., Zurko, M.: Usable Security. IEEE Internet Computing 21 (2017)
16. Yildirim, E.: The importance of information security awareness for the success of business enterprises. AHFE (2016)