Trustworthy Transparency: Only Provably Correct Explanations can Save the World

Maike Schwammberger[1]

Automated and self-learning software systems are increasingly used in a variety of domains and in people's everyday life: from driving assistance systems and products manufactured in smart factories to smart home technologies and applications on our smartphones. Often, the level of automation and system functionality might be known to stakeholders interacting with the automated system to some degree; For instance, an owner of a semi-automated vehicle will have a certain degree of knowledge about the automated distance keeping functionality in their car. However, there still might be features that they do not understand, e.g. the car's behaviour in some special outlier situations (cf. exceptionally bad weather).

We perceive two key reasons why the design and functionality of such automated and self-learning software systems must be made more transparent: (a) all stakeholder groups that interact with such systems need to be sufficiently informed abut the systems' functionality, e.g., to be enabled to safely interact with the system, and (b) with more system transparency, the correct functionality of such systems can be better analysed and verified. We postulate that, without a certain level of system transparency, a system should not be launched into our markets and with that be integrated into our societal contexts. To achieve an increase in system transparency, we develop and investigate a system's self-explainability capabilities. With this, a system is capable of explaining it's decision making process. Our research motivation is that only trustworthy, meaning reliable and correct, explanations can increase system understandability and transparency.

A challenge for engineering self-explainability is the sheer number and complexity of many system's components: Even a (seemingly) simple system like a robot vacuum cleaner comprises a selection of different software components. These could, e.g., be a sensor processing unit for avoiding collisions, a communication unit for interacting with a connected smart home system, a decision making unit for deciding when the robot must return to it's charging station and an AI unit that learns a map about the area that needs cleaning. Due to this level of complexity, having a system engineer write explanations manually cannot be desirable and would certainly lead to human errors. This would then result in unreliable, even incorrect, explanations. For this, we focus on automatically extracting provably correct explanations from system models.

Further on, we investigate different stakeholder groups that receive explanations in different types of situations. With appropriate and situation- and stakeholder-dependent explanations, a system engineer can improve and debug the system during design time, an end-user is enabled to use the system safely and to trust in it's automated decisions, political and societal bodies can decide whether to allow a system to be launched into the markets and lawyers can decide who is to be blamed in tort claim cases that involve automated and self-learning systems. To allow for explanations to be used on such central societal levels, we investigate means to verify correctness, relevance and adequacy of our automatically generated explanations.

[1] Karlsruhe Institute of Technology, Karlsruhe, Germany schwammberger@kit.edu