# Visiting structural differences of explanations

Jakob Michael Schoenborn<sup>1,2</sup> schoenborn@uni-hildesheim.de

 <sup>1</sup> Intelligent Information Systems, University of Hildesheim, Samelsonplatz 1, 31141 Hildesheim, Germany
<sup>2</sup> German Research Center for Artificial Intelligence (DFKI), Trippstadter Str. 12, 67663 Kaiserslautern, Germany

Abstract. During the very recent two years, the rising interest in explanations is undeniable with much more research focusing XAI and conferences using this topic as their main theme. We present our current research in this area and provide further steps we are taking to structure the vast different facets of an explanation (attributes, goals, types, targets, ...). The reason behind this structural effort is to enable an XAI component to effectively find the most appropriate explanation for a given user. We suggest that this explanation can change during the course of a conversation and depends heavily on the user's current emotional state.

Keywords: Explanations  $\cdot$  Structure  $\cdot$  Facets of Explanations

## 1 Research summary: Introduction and motivation

#### 1.1 Past approaches

Last year during the ICCBR-DC-2018 we motivated to lay the foundation to build an explanation-aware case-based reasoning system from scratch to be usable in any domain. The main issue was to decide whether to start from scratch without any current knowledge (since there might be no accessible knowledge, e.g., because of a new domain) or to reduce the current knowledge to distinguish between domain-dependent and domain-independent knowledge. As the received feedback from the reviewers and from the fruitful discussions during the ICCBR-2018 proposed, the best approach is probably to do both and to find a good balance between required amount of knowledge and quality of the issued explanation. Accordingly, we further investigated the common sense definition of an explanation to get an idea how to measure the quality of an explanation.

As the first step, we distinguished between argumentation and explanation at the  $8^{th}$  German Workshop on Experience Management (GWEM19). During our everyday life, we might encounter situations with incomplete information, e.g., passing by two conversing people while switching between university campus and hearing the conversation snippet "... when we are leaving, she usually sleeps ...". Curiously, we try to infer on who is she. Thinking about this situation,

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

most people argue that *she* is some kind of pet, this *can* (but not necessarily) be enhanced by adding the information that two female students are talking to each other. In contrast, if the added information is that two male students are talking to each other, the likelihood that *she* is a pet decreases and increases towards *she* as their girlfriend. However, adding another piece of information, one might argue that, if it is known that these students are studying social work and due to their schedule they are visiting families on a regular basis, *she* could also refer to an elder woman whom they are taking care of. Using this example, we wanted to emphasize on how drastically any piece of information can change the outcome of a reasoning process.

To identify the common sense on the difference between argumentation and explanation, we did a preliminary analysis with 45 participating adults with mostly no IT background (especially no XAI background) whether an argumentation or an explanation is preferred using a small web-based project. Among the multitude of possibilities to define the following two terms, we provide another definition of an argumentation and an explanation<sup>3</sup>:

An argumentation is a reason in which the fact functions as evidence in support of the conclusion. Its goal is to <u>convince</u> the conversational partner on the validity of the conclusion.

An explanation is a supportive, personalized piece of information on top of a provided conclusion. In contrast to an argumentation, its goal is to help the conversational partner in understanding the reasoning behind the conclusion and its outcome.

The user of the website becomes confronted with an image or a situation (see Fig. 1 before reading further). Each of these have two different interpretations - the user has to pick one. After the user has saved the selection, one explanation and one argumentation towards the not chosen interpretation will be given. The user has the option to change the selection or to keep the opinion. If the selection has been changed, the referring argumentation or explanation gains plus 1 score. Closing, the user will be asked about the own opinion. Overall, the insights gained from this experiment:

- Once an opinion has been established, it is hard to change it (but on a bright side: Providing false information to lure the user was not successful)
- The participants do not distinguish between an argumentation and an explanation (or define them completely in contrast to how we did)

Considering the second insight, we tried to formalize  $\operatorname{Arg}_B$  as fact-based and neutral as possible, which has been the most successful option to change a users mind. However,  $\operatorname{Arg}_B$  can also be seen as a Counterfactual Explanation (see below). This supports the arising consensus in the current XAI literature, that argumentation and explanation does not necessarily differentiate from each other, or to be more precise, that the set of possible explanations contains the set of

<sup>&</sup>lt;sup>3</sup> Mostly based on the definition of an argument by Toulmin

	Please choose:	
Y Y Y Y	A) Pillar with a hole in the inside	
	B) Pillar with a pyramid on top	
	If the user chose the hole $(A)$ :	
	$\operatorname{Arg}_B$	Consider the angle of the solar irradiation. If the center
		would be a hole, shouldn't the left, inner side be shaded
		then?
and the share of the	$\operatorname{Exp}_B$	Imagine a pyramid on top of the pillar and place your
		hand upon it. Would you rather change your mind?
	If the user chose the pyramid $(B)$ :	
	$\operatorname{Arg}_A$	80% of surveyed users agreed on the hole. Would you like
		to change your opinion?
	$\operatorname{Exp}_A$	Focus on the brighter surfaces. Could you imagine how
		you place a ball into the hole? If so, would you rather
		change your mind?

**Fig. 1.** One exemplary image of a situation. Here, most people who initially chose option A changed their opinion after following  $\text{Arg}_B$ . When confronting a user with a situation, a textual scenario is given instead of a picture.

possible argumentations. Furthermore, throughout discussing the results, one of the possible next steps is to investigate the possibility of changing an established opinion during the course of a conversation - and how an explanation has to be adjusted to support this.

## 1.2 Current approach

In addition to the listed goals an explanation by Lillehaug [8], Sørmo et al.[11] and the evaluation dimensions for predictions by Leake [6], we are looking to further formalize an explanation using a subset of these approaches. We begin and for the present end here by providing five different type of explanations.

**Case-Based Explanation (CBE)** has largely been presented by Sørmo et al. [11] among others. Case-based reasoning supports the decision making process of a current problem by searching for the most similar case of the concrete past experiences stored in a casebase. This is intuitively very close to actual problem solving in real-life scenarios. To provide the suggested case can be seen as an explanation; optionally paired with the used similarity measure to inform the user why this case has been retrieved as the most similar one. As Sørmo et al. pointed out – and Binns et al. confirmed during their practical field study [1] – Case-Based explanations have their limitations by minimal knowledge requirements of the user about the reasoning process and the own applicability ("Because this happened to x doesn't mean it happens to me as well!", as a participant of the study stated [1]).

**Rule-Based Explanation (RBE)** is usually based on distinct attribute-valuepairs and their respective valid frame of acceptable pairs. These borders are often very sharp, i. e., the similarity switches from one to zero (or vice versa) between two integer values, which can be loosened by using fuzzy logic. Guidotti et al. proposed a solution called LORE (LOcal Rule-Based Explanations) [4]. A local decision is derived from a decision tree, e. g., by following one path, or a random forest and contains a logical rule and a set of counterfactual rules. The logical rule consists of a set of attribute-value pairs  $\{x\}$  of the given domain paired with the proposed decision c(x) while the counterfactual rule  $\Phi$  contains the minimal number of changes in the feature values that would reverse the decision of the predictor. The explanation is the combination of those. The authors provide an example in the credit debt domain [4]:

$$\begin{split} e = \langle r = \{ age \leq 25, job = none, amount > 5k \} \rightarrow deny, \\ \varPhi = \{ (\{ age > 25, amount \leq 5k \} \rightarrow grant), \\ (\{ job = clerk, car = yes \} \rightarrow grant) \} \rangle \end{split}$$

This is an example of the mentioned sharp borders (e. g., age) where the outcome grant/deny can change very quickly. This rule-based approach is very similar to a sensitivity analysis which can (in combination with rules) also be used during the adaptation process in CBR (and thus CBE). Nevertheless, RBE can be treated as a single approach due to its simplicity and can be a very cost-efficient component in an XAI component (provided there is an efficient process to maintain the rulebase).

**Model-Based Explanation (MBE)** depends heavily on the modeler. Models are by definition simplified depictions of the reality which have been developed to fulfill a certain task. Consequently, this results in the observation that there cannot be *the one* model but rather multiple possible valid models. Nevertheless, models are required to issue predictions, e. g., models allow us to imagine and to choose the right present for a person we like. This can also be helpful by providing explanations, as Bokulich describes using an example of the feather coloration of sparrows: " [...] *that allows the sparrows to avoid unnecessary conflicts over resources; dark birds are dominant and displace the pale bird from food sources*" [2, p. 2]. This knowledge retrieved through the model (explanans) can be used and be part of an explanation on why a pale bird does not try to contend the food source (explanandum). MBE are often used to answer "why-questions".

**Emotional-Based Explanation (EBE)** do not follow any rational logic but are rather based on the current emotional state of the user. These often do result into unacceptable explanations, e.g. "because it's like that!" or "because I'm always right!". Nevertheless, whenever we decide to (dis-)agree on an explanation, we evaluate the problem cognitively and take a decision emotionally [9]. These decisions are to some extent biased by environmental conditions such as increased stock returns during sunshine [5] and judges being less likely to condemn the accused person after eating or taking a break [3]. Lerner et al. identified eight major themes of emotional impact on judgment and decision making which further describe the influence, e.g., taking in general the rather safe option instead of a riskier but with higher potential outcome [7]. It can become an important (but admittedly very difficult) aspect to identify when such an EBE arises during the course of a conversation and to make the user aware of his/her current emotional state. We used this kind of explanation for  $\text{Exp}_A$ and  $\text{Exp}_B$  in Fig. 1 by asking the participant to *imagine* another approach on how to perceive the presented image.

**Counterfactual Explanation (CFE)** have been covered partly in MBE. Some authors, when writing about explanations, are distinguishing between why- and why-not explanations. CFE are basically why-not explanations. At first glance it might seem unintuitive why an XAI component should list a range of which are not applicable to a given problem, but exactly this can result in increasing trust of the user to the XAI component. As stated by Sokol et al, CFE fit into dialogues since they are able to correct the user's model of the current domain by narrowing down the acceptable range of certain attribute-value pairs [10].

## References

- Binns, R., Kleek, M.V., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N.: 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. CoRR (2018), http://arxiv.org/abs/1801.10408
- Bokulich, A.: Models and Explanation. Magnani, L. and Bertolotti, T. (eds.) Springer Handbook of Model-Based Science p. 103–118 (2017)
- Danziger, S., Levav, J., Avnaim-Pesso, L.: Extraneous factors in judicial decisions. Proceedings of the National Academy of Sciences of the United States of America 17(108), 6889–6892 (2011)
- Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Gianotti, F.: Local Rule-Based Explanations of Black Box Decision Systems (2018)
- 5. Hirshleifer, D., Shumway, T.: Good Day Sunshine: Stock Returns and the Weather. The Journal of Finance **LVIII**(3) (2003)
- Leake, D.: Goal-based explanation evaluation. Cognitive Science 15, 509–554 (1991)
- Lerner, J.S., Li, Y., Valdesolo, P., Kassam, K.: Emotion and Decision Making. Annual Review of Psychology pp. 799–823 (2015)
- 8. Lillehaug, M.B.: Explanation-aware Case-based Reasoning (2011), master Thesis at the NTNU Trondheim
- Loewenstein, G.F., Weber, E.U., Hsee, C.K., Welch, N.: Risk as Feelings. Psychological Bulletin 127(2), 267–286 (2014)
- Sokol, K., Flach, P.: Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. pp. 5868–5870. International Joint Conferences on Artificial Intelligence Organization (7 2018). https://doi.org/10.24963/ijcai.2018/865, https://doi.org/10.24963/ijcai.2018/865
- Sørmo, F., Cassens, J., Aamodt, A.: Explanation in case-based reasoningperspectives and goals. Artificial Intelligence Review 24(2), 109–143 (Oct 2005), https://doi.org/10.1007/s10462-005-4607-7