From Explanations to Intelligible Explanations[*]

Sylvie Coste-Marquis$^{1\,[0000-0003-4742-4858]}$ and Pierre Marquis$^{1,2\,[0000-0002-7979-6608]}$

[1] CRIL, Univ Artois & CNRS, Lens, France
{coste,marquis}@cril.fr – www.cril.fr
[2] Institut Universitaire de France, France

Abstract. Automatically deriving intelligible explanations to decisions made by an AI system is a challenging task. In this paper, the stress is laid on the intelligibility issue, which concentrates a part of the difficulty of the problem, and relies on the fact that defining what a "good" explanation is does not solely concern what should be explained (the explanandum), but also depends on who receives the corresponding explanans (the explainee). We sketch some general results about intelligibility, that do not rely on specific assumptions on the AI system at hand. A notion of projection is used to characterize among the consequences of an explanation those which can be understood by the user. We evaluate the projection operation in terms of intelligibility, information, and explainability.

Keywords: Explainable AI · Intelligible explanations · Projection.

1 Introduction

Explainability is the degree to which a human being can understand why a decision has been made. It is an important issue, especially when decisions are generated automatically by AI systems, including classifiers and other machine learning (ML) models. Accordingly, there has been a growing body of work on explainable and robust AI (XAI) for the past couple of years.

In this paper, the focus is laid on explanations represented by *logical formulae*. The virtue of logical settings is that a formal meaning can be given to explanations. Obviously enough, there exist many logic-based notions of explanations. For instance, a well-known explanation model (that gave rise to a an abundant literature for centuries) is abduction. Abductive explanations are statements built up from a specific alphabet, and they must entail (at least a part of) the explananda. For other scenarios, a less demanding explanation model can be considered, where explanations are only expected to be consistent with the explananda. Such a less demanding model is considered in model-based diagnosis [9,4], where a diagnosis for a system can be considered as an explanation of the discrepancy between the observed behaviour of the system and its expected one when every component is functioning normally.

[*] Many thanks to the anonymous reviewers for their comments and insights. This work has benefited from the support of the AI Chair EXPEKCTATION of the French National Research Agency (ANR).

Whatever the model, representing explanations as logical formulae is not enough to ensure that they are intelligible. Indeed, even when the explanation under consideration is given as a simple fact, it can be meaningless for the explainee, just because it is totally unrelated to the concepts she/he/it is aware of. In such a case, what can be done with the explanation that has been computed? How to make it somewhat intelligible while preserving as much information as possible? Is it possible to do so without questioning its explanatory power?

As advocated in [3], intelligibility is among the research questions pertaining to XAI that have not been explored in depth, and as such, it should receive more attention. Accordingly, in this paper, the stress is laid on the intelligibility issue, and specifically on the communication problem in explanation, i.e., the fact that the explanation is *for someone* [8]. We consider a simple user model, consisting of a *logical vocabulary* (a set of facts - atomic propositions - which are supposed to be meaningful for the user). We do not commit to any specific AI system (e.g., a classifier). Instead, we assume the existence of a (logic-based) domain theory, from which concepts of explanations can be defined and which can be (tentatively) exploited by the AI system to make the explanations intelligible (or in general ”more intelligible”) once they have been generated. We focus on two concepts of explanations (abductive explanations and consistent explanations). We present a notion of *projection* that can be used to characterize, among the consequences of an explanation, those which can be understood by the explainee, i.e., those that can be expressed using her logical vocabulary. We evaluate the projection operation in terms of intelligibility, information, and explainability.

2 Two Explanation Models

We now present the two concepts of explanations (abductive explanations and consistent explanations) that are considered in the rest of the paper. The following logic-based setting for explanations elaborates a bit over the one presented in [2]. Let $PROP_{PS}$ be a propositional language built up from a finite set PS of symbols and interpreted in a classical way.

Definition 1 (abductive/consistent explanations). *Let T be a propositional formula of $PROP_{PS}$ (a domain theory), that is supposed consistent, A a subset of propositional symbols of PS (the assumptions), M a finite set of propositional formulae of $PROP_{PS}$ (the manifestations) and a subset M^* of it (the conjunctively-interpreted set manifestations to be explained, alias the explananda).*

- *A conjunction γ of variables from A is an* abductive explanation *for M^* w.r.t. T and M if and only if*
  - $\forall m \in M^*, T \wedge \gamma \models m$,
  - *$T \wedge \gamma$ is consistent.*
- *A conjunction γ of variables from A is a* consistent explanation *for M^* w.r.t. T and M if and only if $T \wedge \gamma \wedge M^*$ is consistent.*

The largest M' such that $M^* \subseteq M' \subseteq M$ and γ is an explanation for M' w.r.t. T and M is referred to as the set of manifestations that are *covered* by γ.

In this setting, an (abductive / consistent) explanation must explain all the manifestations for which an explanation is sought (those of M^*), and possibly more. Clearly enough, any abductive explanation is a consistent one, but the converse does not hold. Observe that though explanations are structurally simple (as conjunctions of atoms) in these models, it is not possible in general to guarantee that a single explanation of the manifestations to be explained exists. It can be the case that no explanation can be found, and alternatively, it may happen that many explanations (and sometimes exponentially many) are possible. Preference criteria (e.g., minimality and/or coverage) can be used to restrict the set of candidate explanations, going from explanations to *preferred explanations*. However, the set of preferred explanations is exponentially large in the general case for many preference criteria.

Example 1. Let $T = (ms \Rightarrow (bv \wedge ss \wedge he)) \wedge (my \Rightarrow (bv \wedge \neg he))$ (the meaning given to the atomic propositions occurring in this formula will become clear soon). When $A = \{ms, my, co\}$, $M^* = \{bv\}$ and $M = \{bv, ss\}$, the atoms ms, my, are two (minimal) abductive explanations for M^* w.r.t. T and M. The set of manifestations covered by ms is $\{bv, ss\}$ and the set of manifestations covered by my is $\{bv\}$. Every consistent term over A that does not imply $ms \wedge my$ is a consistent explanation for M^* w.r.t. T and M.

3 Looking for Intelligible Explanations

Explaining is usually a hard task, for a number of reasons. Among them are the number of explanations and the computational complexity of deriving explanations (e.g., computing one explanation is intractable for the two explanation models presented before). Even when explanations are structurally simple, not numerous ... and provided for free, we are not necessarily done. Indeed, it can be the case that the explanations that are reported are useless because they are *not intelligible*.

To make this more precise, let us consider two agents, an explanation provider (or explainer) and an explanation receiver (or explainee). Each of those two agents can be a human being or an artificial agent (the pieces of information exchanged by the two agents can be made formal and their exchange is ruled by protocols that can be automated). The purpose of the explainer is to provide the explainee with intelligible explanations. For the sake of illustration, let us consider the following scenario:

Example 2 (Example 1, cont'd). Abraham goes to her ophthalmologist because he has some eye trouble. Abraham believes that he suffers from myopia. Abraham indicates to her physician that he has a blurred vision. After having examined him, her doctor suspects that Abraham suffers from *Marfan syndrome*. It is the first time that Abraham hears this disease name (this term is totally meaningless for Abraham). Though the fact that Abraham suffers from Marfan syndrome can be considered as an intelligible explanation of the symptoms shown by Abraham from the doctor point of view, it is not from Abraham's point of view since it is entirely unrelated to the concepts Abraham is aware of.

Formally, consider the domain theory T appearing in Example 1 where the variables used have the following meanings:

- ms: “Abraham suffers from Marfan syndrome”.
- my: “Abraham suffers from myopia”.
- bv: “Abraham has a blurred vision”.
- ss: “Abraham has the thumb sign”. The thumb sign (or Steinberg’s sign) is elicited by asking the person to flex the thumb as far as possible and then close the fingers over it. A positive thumb sign is where the entire distal phalanx is visible beyond the ulnar border of the hand, caused by a combination of hypermobility of the thumb as well as a thumb which is longer than usual.
- co: “Abraham suffers from conjunctivitis”.
- he: “Abraham suffers from a hereditary disease”.

The explanation “Marfan syndrome” can be generated automatically as a minimal abductive explanation $\gamma = ms$ for M^* w.r.t. T and M (in the sense of Definition 1), where A, M^*, T and M are as reported in Example 1. The manifestations M^* are explicitly reported by Abraham who asks her physician for an explanation of them. The ms explanation is short, and structurally simple. It is meaningful for the ophthalmologist because she knows the domain theory T, but it is *not intelligible* by Abraham (who probably has an incomplete domain theory since he is not a physician).

3.1 Making an explanation intelligible through projection

The issue is now to determine how to take advantage of the user model, which can be more or less sophisticated, to derive meaningful information from explanations that cannot be understood as such. A very simple abstraction of the explainee is given by her *logical vocabulary*, i.e., the set of atomic propositions that are supposed to be intelligible. Explanations can then be projected onto this vocabulary:

Definition 2 (projecting an explanation onto a vocabulary). *Let γ be a propositional formula of $PROP_{PS}$ (an explanation). Let U be a subset of PS (the user vocabulary). Let T be a propositional formula of $PROP_{PS}$ (a domain theory), that is supposed consistent. The* projection *of γ onto U given T is the set $\Pi(\{\gamma\}, T, U)$ of all logical consequences of $T \wedge \gamma$ belonging to $PROP_U$.*

Example 3 (Example 1, cont’d). The discussion she had with Abraham suggested that Abraham’s vocabulary contains my, bv, he. Hence the physician assumes that $U = \{my, bv, he\}$. Then she may project $\gamma = ms$ onto U given T. The resulting set is equivalent to $bv \wedge \neg my \wedge he$. Doing so, the physician makes γ somewhat intelligible to Abraham, indicating (among other things) that the disease she suspects Abraham suffers from explains the blurred vision symptom, and that unlike myopia, it is a hereditary disease.

By definition, the projection of an explanation onto a vocabulary given a domain theory is an infinite set. In order to make use of it, it is important to associate with it a finite representation that can be computed by an agent (human or artificial), as we did it in the example above. It turns out that computing a finite representation of the projection of the explanation onto a user vocabulary amounts to removing second-order quantifications in a logical formula, which is also known in the propositional case as forgetting propositional variables in a formula. To be more precise, projecting γ onto U given T consists in *forgetting* in $T \wedge \gamma$ every variable that does not belong to U [7,5,1]. Many results about forgetting can be leveraged in this respect.

Observe that the idea of projection considered here is independent of the nature of the explanation. It makes sense as soon as explanations take the form of logical statements. Especially, one can take advantage of it for explanations that are not abductive/consistent explanations.

Note also that replacing the domain theory T by its projection onto the user vocabulary U before computing explanations, or alternatively restricting the set A of assumptions to $A \cap U$, would not have the same effect as projecting explanations onto U given T: doing so would not lead to the same set of explanations in the general case, so that the set of intelligible consequences that could be deduced from an explanation may heavily differ as well.

Example 4 (Example 1, cont'd). The projection of T onto U is equivalent to $my \Rightarrow (bv \wedge \neg he)$, and w.r.t. this projected theory and M, there is only one minimal abductive explanation for M^*, namely my. Similarly, assuming that A has been reduced to $A \cap U = \{my\}$, my is the unique minimal abductive explanation for M^* w.r.t. T and M.

Clearly enough, unlike ms, my does not cover the manifestation ss and for this reason, it has been considered as less preferred. Finally, my has consequences over U given $my \Rightarrow (bv \wedge \neg he)$ that conflict with the consequences of ms over U given T since the former is not a hereditary disease ($\neg he$ is a consequence of my given $my \Rightarrow (bv \wedge \neg he)$) while the latter is a hereditary disease (he is a consequence of ms given T).

3.2 What is got and what is lost when projecting an explanation

Obviously, replacing an explanation by its projection onto a user vocabulary given a domain theory is not a neutral operation in general. Thus, it is important to evaluate the projection operation in terms of intelligibility, information, and explainability.

First of all, projecting an explanation onto a user vocabulary can *only increase the amount of intelligible information* furnished to the user, assuming that the user has her/his/its own knowledge base T_U (a propositional formula) such that $U = Var(T_U)$, and $T \models T_U$ (this means that the explainee has possibly a partial knowledge of the domain theory of the explainer, but has no wrong beliefs). Especially, whenever a representation of the projection of an explanation γ onto U given T is provided to a user, she can derive thanks to it and using her restricted domain theory T_U the same set of consequences over U as if she was fully aware of the domain theory T:

Proposition 1. *Let γ, T, T_U be three formulae from $PROP_{PS}$ such that $T \models T_U$, and let $U \subseteq PS$. The set of logical consequences over U of $T_U \wedge \gamma$ (i.e., the information that can be deduced by the user when γ is added to her knowledge base) is a subset of the set of logical consequences over U of $\{T_U\} \cup \Pi(\{\gamma\}, T, U)$, which coincides with $\Pi(\{\gamma\}, T, U)$; using symbols:*

$$\Pi(\{\gamma\}, T_U, U) \subseteq \Pi(\Pi(\{\gamma\}, T, U), T_U, U) = \Pi(\{\gamma\}, T, U).$$

However, the projection process leads to an *information loss* in the general case, meaning that the projection of γ onto U given T is not equivalent to $T \wedge \gamma$ in the general case, but is "only" a logical consequence of it:

Proposition 2. *Let γ, T be two formulae from $PROP_{PS}$ and let $U \subseteq PS$. We have $T \wedge \gamma \models \mathit{\Pi}(\{\gamma\}, T, U)$ but in the general case we do* **not** *have $T \wedge \gamma \equiv \mathit{\Pi}(\{\gamma\}, T, U)$.*

Indeed, the projection of an explanation onto a vocabulary does not necessarily correspond to an explanation itself. In fact, this depends on the explanation model at hand. Thus, in the abductive model, an explainability loss may occur:

Example 5 (Example 1, cont'd). As explained previously, in order to explain the manifestations that are observed, the physician prefers the abductive explanation ms to the abductive explanation my because ms covers more symptoms than my. The projection of ms onto U is equivalent to $bv \wedge \neg my \wedge he$ but this formula cannot be considered as an abductive explanation for M^* since it is not a conjunction of assumptions from A. Furthermore, the only conjunction of variables from $A \cap U$ that is consistent with it is the empty conjunction. This empty assumption is consistent with T but it does not explain the manifestations M^* (we have $T \not\models bv$).

Contrastingly, consistent explainability is preserved though projection, simply because this operation is consistency-preserving (for any γ over A such that $T \wedge \gamma \wedge M^*$ is consistent, $\mathit{\Pi}(\{\gamma\}, T, U) \cup \{T\} \cup M^*$ is consistent).

In a longer version of this research note (see www.cril.fr/~marquis/intelligible.pdf), we focus on the case when the explanation under consideration can be *reformulated* using the user vocabulary in the domain theory, so that no information is actually lost when the explanation itself is replaced by an equivalent reformulation (this amounts to a *definability* issue [6]). We also explain how to *compute and simplify projections* using theory reasoning.

References

1. Delgrande, J.P.: A knowledge level account of forgetting. J. Artif. Intell. Res. **60**, 1165–1213 (2017)
2. Eiter, T., Gottlob, G.: The complexity of logic-based abduction. J. ACM **42**(1), 3–42 (1995)
3. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 93:1–93:42 (2019)
4. de Kleer, J., Mackworth, A.K., Reiter, R.: Characterizing diagnoses and systems. Artificial Intelligence **56**, 197–222 (1992)
5. Lang, J., Liberatore, P., Marquis, P.: Propositional independence: Formula-variable independence and forgetting. J. Artif. Intell. Res. **18**, 391–443 (2003). https://doi.org/10.1613/jair.1113
6. Lang, J., Marquis, P.: On propositional definability. Artif. Intell. **172**(8-9), 991–1017 (2008). https://doi.org/10.1016/j.artint.2007.12.003
7. Lin, F., Reiter, R.: Forget it! In: AAAI Fall Symposium on Relevance. pp. 154–159 (1994)
8. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
9. Reiter, R.: A theory of diagnosis from first principles. Artificial Intelligence **32**, 57–95 (1987)