From Explanations to Intelligible Explanations (Extended Version)*

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 many cases. In this report, 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. Obviously enough, in general, the trace of an algorithm cannot be considered as an explanation: though it justifies why the output has been generated from the input, such a trace is not comprehensible most of the time. In the past decade, ML techniques have revolutionized vision, speech, language understanding, and many other fields. However, the most powerful ML models in term of quality of predictions are still poorly explainable.

In the meanwhile, the explanation requirement for decisions based on automated processing had become a legal issue in Europe since the implementation of the General Data Protection Regulation (EU) 2016/679 ("GDPR") on May 25th, 2018 [12]. GDPR is a regulation in EU law on data protection and privacy for all individual citizens of the European Union (EU) and the European Economic Area (EEA), see also [12]. GDPR stipulates (Recital 71) that: "The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him

* This work has benefited from the support of the AI Chair EXPEKCTATION of the French National Research Agency (ANR).

or her which is based solely on automated processing [...] In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision."

Accordingly, there has been a growing body of work on explainable and robust AI (XAI) for the past couple of years (see among many other references [3,11,16,15,2,20,26,29,30]).

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, so that any reasoning process based on those explanations can be analyzed (for instance, to determine whether or not it is truth-preserving). Obviously enough, there exist many logic-based notions of explanations. Accordingly, a number of formal settings have been designed to characterize explanations in logical terms and to reason about them.

Among others, abduction (often defined as inference to the best explanation) gave rise to a number of formal developments for centuries (it was already considered by Aristotle in his "Prior Analytics"), and to an abundant literature in philosophy and in AI. The basic pattern of abductive inference can be exemplified as follows. Suppose that I want to explain why Socrates is mortal. Knowing that every man is mortal, an explanation is that Socrates is a man. Clearly, abduction is ampliative, meaning that the conclusion that is reached goes beyond what is (logically) contained in the premises, thus it can be wrong. To illustrate it and quote Ionesco in his drama "Rhinoceros": Socrates is mortal, every cat is mortal, therefore Socrates is a cat.

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 [28,17], 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.

Whatever the model, representing explanations as logical formulae is not enough to ensure that they are comprehensible. Especially, it cannot be guaranteed that the explanation receiver (here, a human being) will be able to draw simple reasonings from the explanations that have been provided if the corresponding formulae are not of small size, if they have a complex structure or if they are too numerous to be embraced as a whole. But even when there is a single explanation 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 [14], intelligibility is among the research questions pertaining to XAI that have not been explored in depth, and as such, it should receive more attention. To make a step in this direction, in the following, the stress is laid on the intelligibility issue about explanations. We specifically focus on the communication problem in explanation, i.e., the fact that the explanation is *for someone* [25]. 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). Based on it, our purpose is to address the research questions listed above.

Our investigation thus departs from many recent works that focus on deriving explanations (of various kinds) for specific AI systems (e.g., a classifier) and typically aim to explain the output returned by the system (e.g., the prediction done) from the corresponding input, by synthesizing the trace of the computation. Especially, we do not commit to any specific AI system. Instead, we assume the existence of a (logic-based) domain theory, from which concepts of explanations can be defined and which can be exploited by the AI system to make the explanations intelligible (or in general "more intelligible") once they have been generated. We consider 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. We focus on the specific case of definable explanations. We also explain how projections can be computed and simplified provided that the explanation provider is aware of some part of the knowledge of the explanation receiver (alias the user). We show how the notions of forgetting and definability (which are well-studied concepts in logic – forgetting goes back to George Boole) can be exploited to reason about the intelligibility of explanations.

The rest of the paper is organized as follows. After some formal preliminaries (Section 2), we present two abductive model for explanations in Section 3. In Section 4, we define the notion of projection, study some of its properties, and explain how to compute and simplify projections. Finally, Section 5 concludes the paper and presents some perspectives for further research.

2 Formal Preliminaries

$PROP_{PS}$ denotes the propositional language built up from a finite set PS of symbols, the Boolean constants $\top$ (true) and $\bot$ (false), and the connectives $\neg$, $\vee$, $\wedge$, $\Rightarrow$, $\Leftrightarrow$. $Var(\phi)$ denote the set of propositional variables occurring in the formula ϕ. If $X \subseteq PS$, $\bar{X}$ denotes the subset of PS given by $PS \setminus X$.

A clause is a finite disjunction of literals. A `CNF` formula is a conjunction of clauses. A term is a finite conjunction of literals. A canonical term over a subset X of PS is a consistent term in which every variable $x \in X$ occurs either as a positive literal (x) or as a negative literal ($\neg x$). A `DNF` formula is a disjunction of terms.

An interpretation ω is an assignment of a truth value to each variable of PS. Formulas are interpreted in the classical way. $\models$ denotes entailment and $\equiv$ denotes logical equivalence. Every finite set of formulas is interpreted conjunctively. An implicant of a formula ϕ is a term γ such that $\gamma \models \phi$. Any formula is equivalent to the disjunction of its implicants (when terms are considered up to logical equivalence, any formula has finitely many implicants).

3 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 [8].

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 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*.

Most plausible explanations are typically preferred for an obvious reason. However, it is not always easy to characterize such most plausible explanations due to a lack of plausibility information. Thus, when considering a logic-based setting for representing explanations, *minimal* explanations are often considered, i.e., explanations that are as weak as possible from a logical standpoint. Assuming that a probability distribution over the set all explanations exist (but is unknown), a first explanation that is a logical consequence of a second explanation is at least as probable as the latter. Especially, focusing on minimal explanations ensures that explanations do not contain pieces of information that are irrelevant to the explanandum.

In some cases, several criteria must be aggregated in order to define a notion of preferred explanation. Thus, in the setting for explanations described above, a trade-off can be looked for: a first explanation can be considered as preferred to a second explanation when the former covers a superset of manifestations covered by the latter, and this comparison relation typically conflicts with the minimality criterion (more assumptions are often necessary to cover more manifestations). The two comparison relations (modeled here as partial preorders) can be combined in various ways (a simple one is

lexicographic aggregation: prefer first the explanations covering the maximal (w.r.t. $\subseteq$) subsets of manifestations, and among them, select those which are minimal).

While, by definition, the set of preferred explanations cannot be larger than the set of all explanations, it can still be exponentially large in the size of the input (this happens for abductive/consistent explanations). Furthermore, preferred explanations can be structurally more complex (hence less intelligible) and/or harder to compute than other explanations. Simplicity (Occam's razor, which states that from two explanations the simpler explanation is preferable) is a key criterion that is often considered.

Simplicity is also valuable when intelligibility is expected. In some settings, especially the one for abductive/consistent explanations presented above, simplicity amounts to minimality (a conjunction of variables is a simple – and as logically weak – as it contains few variables).

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 $\{bv\}$. co is irrelevant to the explanandum (this symbol does not even occur in the domain theory T). Every consistent term over A that does not imply $ms \wedge my$ is a consistent explanation for M^* w.r.t. T and M.

Obviously enough, both the structure of the explanations, their sizes and their number impact their comprehensibility. Another dimension in the complexity of explaining is the *computational effort* that must be spent to derive explanations (i.e., all of them, or only one of them, or even to decide whether an explanation exists – this last issue gives a lower bound on the complexity of the other problems: when it is intractable, the problem of generating one / all explanations are intractable as well). Of course, the computational complexity of deriving explanations heavily depends both on their size and structure, as well as of their numbers (when the goal is to compute all of them). However, the size, structure and number of explanations are not the sole parameters that have an impact on the difficulty to derive explanations. The type of explanation under consideration plays also a key role.

Thus, in the consistent explanation model, the problem of deciding whether a consistent explanation for M^* w.r.t. T and M given A exists amounts to deciding whether $T \wedge M^*$ is consistent, which is intractable, but "only" NP-complete. In the abductive model, as easy consequences of results reported in [8], the problem of deciding whether an abductive explanation for M^* w.r.t. T and M given A exists is likely to be more difficult (it is Σ_2^p-complete).

A further aspect to be taken into account when dealing with explanations is the *meta-explanation* issue and its complexity. Explaining is finding out explanations, while meta-explaining is explaining why a given piece of information is an explanation, or, in the case when counterfactual explanations are sought for, why a given piece of information is not an explanation. When explanations are logical formulae from a given language, meta-explanations are objects from the associated meta-language. From a computational complexity point of view, such objects are certificates. Again, the nature

of the meta-explanations and the complexity to derive them depend on the underlying explanation model, and as such they may differ. In the abductive model for explanations, in order to explain why a given γ is an abductive explanation of M^* w.r.t. T and M, one must find a model of T that is compatible with γ (again a short certificate) but also a *proof* of the fact that $T \wedge \gamma \models M^*$ (or equivalently that $(T \wedge \gamma) \Rightarrow M^*$ is a valid formula). Such proofs are not of polynomial size in general for existing proof systems (e.g., the ones based on resolution) and it is unlikely that polynomial-sized proofs of such formulae may exist in any formal system for propositional logic that is sound and complete for the validity question (the existence of such proofs would imply that NP = coNP, that is considered unlikely in complexity theory). Contrastingly, in order to explain why a given γ is a consistent explanation for M^* w.r.t. T and M, one must find a model of $T \wedge M^*$ that is compatible with γ. Such certificates are small (their sizes are equal to the numbers of propositional variables appearing in the input).

Finally, even when the explanations are structurally simple, of small size, not numerous ... and provided for free, we are not necessarily done. Indeed, it can be the case that the explanations that are reported are totally useless because they are *not intelligible*.

4 Looking for Intelligible Explanations

To make the intelligibility issue 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: distant objects are blurry while close objects appear normal for him. Abraham believes that he suffers from myopia, so that eyeglasses will be enough to treat the problem. 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. At that stage what is very important for Abraham is to get all the information he may understand (given her own vocabulary and background knowledge) that are about this disease, especially as to its treatment and prognosis. Is it a severe disease? What are its causes? What are its most serious complications? How to cure it? Abraham is specifically interested in knowing whether his children might suffer from this disease at some point as well.

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 other manifestations from $M \setminus M^*$ (here ss) are observed directly by the doctor (but in the general case, the patient could be aware of them as well). 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).

4.1 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.

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 or 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).

4.2 From projecting to forgetting

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, where the operation of forgetting is defined as follows (see [21,18,6] for more details):

Definition 3 (forgetting). *Let ϕ be a formula from $PROP_{PS}$ and $X \subseteq PS$. The* forgetting *of X in ϕ, noted $\exists X.\phi$, is a quantified Boolean formula over PS, equivalent to a formula from $PROP_{PS}$ that can be inductively defined as follows:*

- $\exists \emptyset.\phi \equiv \phi$;
- $\exists \{x\}.\phi \equiv \phi_{x \leftarrow 0} \vee \phi_{x \leftarrow 1}$;
- $\exists (\{x\} \cup X).\phi \equiv \exists V.(\exists \{x\}.\phi)$.

$\exists X.\phi$ represents the logically strongest consequence ψ (unique up to logical equivalence) of ϕ that is *independent* of X, where ψ is independent of X means that there exists a formula χ from $PROP_{PS}$ s.t. $\psi \equiv \chi$ and $Var(\chi) \cap X = \emptyset$.

Accordingly, forgetting a set of variables within a formula leads to weaken it. To be more precise, if $V \subseteq W$ holds, then $\exists V.\phi \models \exists W.\phi$ holds. Moreover, ϕ is consistent iff $\exists Var(\phi).\phi$ is valid (see [18]).

Many characterizations of forgetting, together with complexity results, are reported in [18]. Noticeably, for every $V \subseteq PS$ and every formula ϕ from $PROP_{PS}$, we have $\exists V.\phi \equiv \exists V_\phi.\phi$, where $V_\phi = V \cap Var(\phi)$ – which means that forgetting variables that do not appear in a formula does not have any effect.

4.3 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:*

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

Proof. $\Pi(\Pi(\{\gamma\}, T, U), T_U, U)$ is equivalent to $\exists \bar{U}.((\exists \bar{U}.(\gamma \wedge T)) \wedge T_U)$. $\exists \bar{U}.((\exists \bar{U}.(\gamma \wedge T)) \wedge T_U)$ is equivalent to $(\exists \bar{U}.(\gamma \wedge T)) \wedge (\exists \bar{U}. \wedge T_U)$ since $\exists \bar{U}.(\gamma \wedge T)$ is independent on $\bar{U}$. Now, when $T \models T_U$, we also have $\gamma \wedge T \models T_U$ so that $\exists \bar{U}.(\gamma \wedge T)$ implies $\exists \bar{U}.T_U)$, showing that $\Pi(\Pi(\{\gamma\}, T, U), T_U, U) = \Pi(\{\gamma\}, T, U)$. Finally, since $\Pi(\{\gamma\}, T_U, U)$ is equivalent to $\exists \bar{U}.(\gamma \wedge T_U)$, and since $\gamma \wedge T \models \gamma \wedge T_U$ when $T \models T_U$, we also have that $\exists \bar{U}.(\gamma \wedge T)$ implies $\exists \bar{U}.(\gamma \wedge T_U)$, showing that $\Pi(\{\gamma\}, T_U, U) \subseteq \Pi(\{\gamma\}, T, U)$, and this completes the proof.

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 $T \wedge \gamma \models \Pi(\{\gamma\}, T, U)$ but in the general case we do* **not** *have $T \wedge \gamma \equiv \Pi(\{\gamma\}, T, U)$.*

Proof. On the one hand, since $\Pi(\{\gamma\}, T, U)$ is equivalent to $\exists \bar{U}.(\gamma \wedge T)$, showing that $T \wedge \gamma \models \Pi(\{\gamma\}, T, U)$ amounts to proving that $T \wedge \gamma \models \exists \bar{U}.(\gamma \wedge T)$, which is obvious since the consequences of $\gamma \wedge T$ over U are straightforwardly among the consequences of $\gamma \wedge T$. On the other hand, the running example shows that $T \wedge \gamma \not\equiv \Pi(\{\gamma\}, T, U)$: from the projection of γ onto his own vocabulary given T, Abraham cannot derive that he suffers from Marfan disease.

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 discussed 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, $\Pi(\{\gamma\}, T, U) \cup \{T\} \cup M^*$ is consistent).

4.4 The case of definable explanations

Now, an interesting question is to determine whether a loss of information actually occurs when a projection is achieved. When some concepts used in the explanation at hand do not belong to the user vocabulary (as it is the case in the running example), one can conclude that some pieces of information have disappeared. More generally, as discussed before, some information are lost whenever $T \wedge \gamma \not\equiv \Pi(\{\gamma\}, T, U)$. However, a less demanding interpretation of information loss also makes sense: it can be the case that 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.

As a matter of illustration, let us consider the sequel of the discussion between Abraham and her doctor:

Example 6 (Example 1, cont'd). At that stage of the consultation, once the physician told Abraham that he suspected that Abraham suffers from Marfan disease, Abraham asks her for a counterfactual explanation: why not considering myopia[3] as an explanation? The doctor then explains that Abraham also has the thumb sign, and myopia does not explain it. Since ss does not belong to U, once again, this explanation is not intelligible by Abraham.

Suppose that the domain theory T contains also the piece of knowledge $ss \Leftrightarrow (ht \wedge lt)$ where ht means that Abraham's thumb is hypermobile and lt means that Abraham's thumb is longer than usual. Given that ht and lt are simple concepts, the physician may assume that Abraham is able to understand them, so that $U = \{my, bv, he, ht, lt\}$.

This time, unlike what happened for ms, the explanation ss that is not intelligible by Abraham can be reformulated using Abraham's vocabulary: ss *precisely* means that Abraham's thumb is hypermobile (ht) and longer than usual (lt).

Deciding whether such a reformulation exists and, if so, computing a representation of it, amounts to a *definability* issue:

[3] my indeed is an abductive explanation in this case, however it is not a preferred one.

Definition 4 (definable explanation). *Let $\gamma, T \in PROP_{PS}$, $U \subseteq PS$. The explanation γ can be considered as definable in terms of the user vocabulary U in the domain theory T whenever there exists a formula $\Phi_U \in PROP_U$ such that γ is equivalent to Φ_U in T, i.e., we have $T \models \gamma \Leftrightarrow \Phi_U$. When γ is definable, any admissible Φ_U is referred to as a definition of γ on U in T.*

When γ is definable in terms of U in T, one can project Φ_U onto U given T instead of projecting γ onto U given T. Indeed, we have

$$\mathit{\Pi}(\{\gamma\}, T, U) = \mathit{\Pi}(\{\Phi_U\}, T, U).$$

This is particularly helpful when the user knowledge base T_U is known to be equivalent to $\exists \bar{U}.T$ since in this situation, instead of providing $\mathit{\Pi}(\{\gamma\}, T, U)$ to the user, one can simply let her known as an explanation that Φ_U holds, and from it, she will be able to deduce every piece of information conveyed by $\mathit{\Pi}(\{\gamma\}, T, U)$.

The definability issue has been considered in logic for decades, focusing on the case when γ is atomic (i.e., in the propositional case, a variable) (see [19] for details):

Definition 5 (explicit definability). *Let $\phi \in PROP_{PS}$, $X \subseteq PS$ and $y \in PS$. ϕ* explicitly defines *y in terms of X iff there exists a formula $\Phi_X \in PROP_X$ s.t. $\phi \models \Phi_X \Leftrightarrow y$.*

Definability and forgetting are strongly connected. Indeed, whenever y is defined in terms of X in ϕ, the definitions of y on X in ϕ are precisely the formulae Φ_X satisfying the following condition (see Theorem 8 in [18]):

$$\exists \bar{X}.(T \wedge y) \models \Phi_X \models \neg \exists \bar{X}.(T \wedge \neg y).$$

Now, it is quite easy to lift the case when γ is a propositional variable to the general case when γ is any formula. To do so, it is enough to prove that a formula γ is definable in terms of U w.r.t. a domain theory T if and only if $T \wedge (x_\gamma \Leftrightarrow \gamma)$ defines the (fresh) variable x_γ (i.e., not belonging to $Var(T) \cup Var(\gamma) \cup U$) in terms of U [19]. Obviously enough, any definition of x_γ on U in $T \wedge (x_\gamma \Leftrightarrow \gamma)$ also is a definition of γ on U in T.

Interestingly, in the abductive model for explanation, the projection of an explanation γ onto a vocabulary U given a domain theory T does not lead to an explainability loss whenever γ is definable in terms of U w.r.t. T:

Proposition 3. *Let $\gamma, T \in PROP_{PS}$, $A, U \subseteq PS$ such that $U \subseteq A$. Let M^*, M be finite sets of propositional formulae of $PROP_{PS}$ such that $M^* \subseteq M$. Suppose that γ is an abductive explanation for M^* w.r.t. T and M and that γ is definable in terms of U in T, so that there exists a formula Φ_U from $PROP_{PS}$ that is a definition of γ on U in T. Let γ_U be any implicant of Φ_U that is consistent with T. Then γ_U is an abductive explanation for M^* w.r.t. T and M.*

Proof. We first prove that Φ_U has an implicant γ_U that is consistent with T. Towards a contradiction, since Φ_U is equivalent to the disjunction of its implicants, if every implicant γ_U of Φ_U is inconsistent with T, then $T \wedge \Phi_U$ is inconsistent as well. But since $T \models (\gamma \Leftrightarrow \Phi_U)$, if $T \wedge \Phi_U$ is inconsistent then $T \wedge \gamma$ also is inconsistent. This

contradicts the fact that γ is an abductive explanation for M^* w.r.t. T and M. Let now consider any implicant γ_U of Φ_U that is consistent with T. It remains to show that $T \wedge \gamma_U \models M^*$. Since $\gamma_U \models \Phi_U$, we have $T \wedge \gamma_U \models T \wedge \Phi_U$. Since $T \models (\gamma \Leftrightarrow \Phi_U)$, we also have $T \wedge \Phi_U \models T \wedge \gamma$. Altogether we get that $T \wedge \gamma_U \models T \wedge \gamma$. Hence if $T \wedge \gamma \models M^*$ then we also have $T \wedge \gamma_U \models M^*$, which completes the proof.

4.5 Computing and simplifying projections

In the general case, the projection of an explanation onto a vocabulary given a domain theory cannot be represented compactly (under standard assumptions of complexity theory, any circuit encoding it is of size exponential in the input size in the worst case, see [18]). This is still the case when the explanation is structurally simple (for instance, when it takes the form of a conjunction of variables, as in the explanation setting considered in this report) and when it is definable in terms of the user vocabulary in the domain theory. This shows that the conditions to be satisfied for considering an explanation as intelligible are not independent one another: projecting an explanation of small size and simple structure onto a user vocabulary to make it intelligible may lead to a projection which is hard to be understood, this time because of its size.

When $\gamma \wedge T$ is given as a CNF formula $\bigwedge_{i=1}^{k} \delta_i$ and $\bar{U} = \{x\} \cup X$, a CNF formula equivalent to $\Pi(\{\gamma\}, T, U)$ can be computed in a recursive way by eliminating x in $\gamma \wedge T$, obtaining thus a new CNF formula equivalent to $\exists\{x\}.(\gamma \wedge T)$, in which the variables of X are then eliminated. Eliminating x in $\gamma \wedge T$ basically amounts to applying the resolution principle: $\exists\{x\}.(\gamma \wedge T)$ is equivalent to the CNF formula consisting of the clauses δ_i of $\gamma \wedge T$ such that $x \notin Var(\delta_i)$, conjoined with all the resolvents on x of the clauses of $\gamma \wedge T$. In the general case, the resulting CNF formula can be of size exponential in the size of $\gamma \wedge T$.

Nevertheless, there exist restrictions under which the projection of an explanation onto a vocabulary given a domain theory can be represented in space polynomial in the size of the input, and can even be computed in polynomial time from it. Thus, when γ is a conjunction of literals and T is a DNNF representation [4], one can compute in polynomial time a representation of the projection as a DNNF representation. This also holds when T is a DNF representation (in that case, the projection will be represented as a DNF formula). This mainly comes from the fact that DNNF and DNF as representation languages offer a conditioning transformation and a forgetting transformation in polynomial time [5]. When the size of U is considered as bounded, the size of the projection remains small as well.

Deciding whether an explanation γ is definable in terms of a vocabulary U given a theory T is "mildly hard" (coNP-complete in general) since it is not necessary to guess a corresponding definition of γ on U in T to solve this decision problem (this comes from the equivalence between explicit and implicit definability and a method to decide the latter [27]). The generation of a definition is typically much more expensive, since under standard assumptions of complexity theory, there is no polynomial-size circuit representing such a definition in the general case [19]. The other way around, when the explanation γ under consideration is a formula that is definable in terms on U in T, it can be the case that a definition of it on U in T that is exponentially smaller than γ

exists. Such a size shift (both ways) between an explanation and a definition of it must be taken into account when the goal is to get an intelligible explanation.

On the other hand, assuming that the explainer is aware of a part T_U^p of the knowledge base T_U of the explainee (alias the user), the explainer can exploit it to *simplify* the projection $\Pi(\{\gamma\}, T, U)$ into a formula $simp(\Pi(\{\gamma\}, T, U), T_U^p)$ using *theory reasoning* w.r.t. T_U^p. More formally, let T_U^p be a formula such that $T \models T_U \models T_U^p$ and $Var(T_U^p) = U$, the objective is to generate a formula $simp(\Pi(\{\gamma\}, T, U), T_U^p)$ satisfying $T_U^p \wedge simp(\Pi(\{\gamma\}, T, U), T_U^p) \equiv \Pi(\{\gamma\}, T, U)$ that is as simple as possible, so as to improve the intelligibility. Indeed, given that the user knows T_U^p, from $simp(\Pi(\{\gamma\}, T, U), T_U^p)$, she will be able to recover using her knowledge base T_U (which is at least as strong as T_U^p) all the consequences of $\Pi(\{\gamma\}, T, U)$ that are not consequences of $simp(\Pi(\{\gamma\}, T, U), T_U^p)$. Especially, doing so, the pieces of information from $\Pi(\{\gamma\}, T_U^p, U)$ that are irrelevant to the explanation γ but appear in $\Pi(\{\gamma\}, T, U)$ because they are consequences of T_U^p are filtered out. Thus, a strong violation of Grice's maxim of quantity [13] is avoided (Grice's maxims are norms governing cooperative communication among agents; the quantity maxim indicates that the contribution of the provider should not contain information that the receiver is already aware of).

Going a step further requires to make precise what "simple" means. This is clearly dependent on the representation chosen for $\Pi(\{\gamma\}, T, U)$. A standard format is the `CNF` one: each conjunct (a clause) is simple enough and the way they are connected is clear as well. If $\Pi(\{\gamma\}, T, U)$ is given as a `CNF` formula, a candidate for $simp(\Pi(\{\gamma\}, T, U), T_U^p)$ is the conjunction of the *theory prime implicates* of $\Pi(\{\gamma\}, T, U)$ w.r.t. T_U^p. Basically, whenever two clauses δ_1, δ_2 of $\Pi(\{\gamma\}, T, U)$ are such that $\delta_1 \wedge T_U^p \models \delta_2$, δ_2 can be removed; furthermore, if a subclause δ_3 of δ_1 is such that $\delta_3 \wedge T_U^p \models \delta_1$, then δ_1 can be replaced by δ_3 (see [23,24] for more details).

5 Conclusion

This paper is centered on the intelligibility question for explanations. A simple user model, consisting of a logical vocabulary (a set of facts which are meaningful for the explainee), has been considered. On this ground, we have presented 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 vocabulary. We have evaluated the projection operation in terms of intelligibility, information, and explainability. We have studied the specific case of definable explanations. We have also sketched how projections can be computed and simplified.

This work calls for many perspectives for further research. One of them is to consider more expressive settings than classical propositional logic and to investigate the extent to which the results presented in the paper can be lifted. Interestingly, the key operation of forgetting has been studied in many logical settings, especially logic programming, modal logics, description logics, and it already gave rise to many papers and some pieces of software (see among others [22,36,37,33,34,7,9,1,35,32,31]). Another perspective is to instantiate our approach to other explanation settings, especially settings for which explanations are event-based (obviously enough, there exist numerous logical settings for modeling and reasoning about actions; furthermore the notion of for-

getting actions has already been considered so far [10]). Finally, more work is needed to figure out the interplay between the criteria that are expected to be fulfilled by intelligible explanations. When the size and/or the structure of a projection that has been simplified and compressed remain(s) too complex, it would make sense to approximate it so as to improve its intelligibility. However, this may have an impact on the explanatory power of the explanation (just as it happens when considering the projection of the explanation instead of the explanation itself since $\Pi(\{\gamma\}, T, U)$ is an approximation of $T \wedge \gamma$, logically speaking the best possible one onto U). It is likely that some trade-offs should be looked for.

Acknowledgments Many thanks to the anonymous reviewers of the XLoKR'20 workshop (see https://lat.inf.tu-dresden.de/XLoKR20/) for their comments and insights on a short version of this report.

References

1. Antoniou, G., Eiter, T., Wang, K.: Forgetting for defeasible logic. In: LPAR'18. pp. 77–91 (2012)
2. Bunel, R., Turkaslan, I., Torr, P.H.S., Kohli, P., Mudigonda, P.K.: A unified view of piecewise linear neural network verification. In: NeurIPS'18. pp. 4795–4804 (2018)
3. Cheng, C., Diehl, F., Hinz, G., Hamza, Y., Nührenberg, G., Rickert, M., Ruess, H., Truong-Le, M.: Neural networks for safety-critical applications - challenges, experiments and perspectives. In: DATE'18. pp. 1005–1006 (2018)
4. Darwiche, A.: Decomposable negation normal form. J. ACM **48**(4), 608–647 (2001). https://doi.org/10.1145/502090.502091
5. Darwiche, A., Marquis, P.: A knowledge compilation map. J. Artif. Intell. Res. **17**, 229–264 (2002). https://doi.org/10.1613/jair.989
6. Delgrande, J.P.: A knowledge level account of forgetting. J. Artif. Intell. Res. **60**, 1165–1213 (2017)
7. van Ditmarsch, H., Herzig, A., Lang, J., Marquis, P.: Introspective forgetting. Synthese **169**(2), 405–423 (2009)
8. Eiter, T., Gottlob, G.: The complexity of logic-based abduction. J. ACM **42**(1), 3–42 (1995)
9. Eiter, T., Wang, K.: Semantic forgetting in answer set programming. Artif. Intell. **172**(14), 1644–1672 (2008)
10. Erdem, E., Ferraris, P.: Forgetting actions in domain descriptions. In: AAAI'07. pp. 409–414 (2007)
11. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. In: CVPR'18. pp. 1625–1634 (2018)
12. Goodman, B., Flaxman, S.R.: European union regulations on algorithmic decision-making and a "right to explanation". AI Magazine **38**(3), 50–57 (2017)
13. Grice, P.H.: Studies in the Way of Words. Harvard University Press (1989)
14. 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)
15. Ignatiev, A., Narodytska, N., Marques-Silva, J.: Abduction-based explanations for machine learning models. In: AAAI'19. pp. 1511–1519 (2019)

16. Katz, G., Huang, D.A., Ibeling, D., Julian, K., Lazarus, C., Lim, R., Shah, P., Thakoor, S., Wu, H., Zeljic, A., Dill, D.L., Kochenderfer, M.J., Barrett, C.W.: The marabou framework for verification and analysis of deep neural networks. In: CAV'19. pp. 443–452 (2019)
17. de Kleer, J., Mackworth, A.K., Reiter, R.: Characterizing diagnoses and systems. Artificial Intelligence **56**, 197–222 (1992)
18. 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
19. 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
20. Leofante, F., Narodytska, N., Pulina, L., Tacchella, A.: Automated verification of neural networks: Advances, challenges and perspectives. CoRR **abs/1805.09938** (2018)
21. Lin, F., Reiter, R.: Forget it! In: AAAI Fall Symposium on Relevance. pp. 154–159 (1994)
22. Lutz, C., Wolter, F.: Foundations for uniform interpolation and forgetting in expressive description logics. In: IJCAI'11. pp. 989–995 (2011)
23. Marquis, P.: Knowledge compilation using theory prime implicates. In: IJCAI'95. pp. 837–843 (1995)
24. Marquis, P., Sadaoui, S.: A new algorithm for computing theory prime implicates compilations. In: AAAI'96. pp. 504–509 (1996)
25. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
26. Molnar, C., Casalicchio, G., Bischl, B.: iml: An R package for interpretable machine learning. J. Open Source Software **3**(26), 786 (2018)
27. Padoa, A.: Essai d'une théorie algébrique des nombres entiers, précédé d'une introduction logique à une théorie déductive quelconque. In: Bibliothèque du Congrès International de Philosophie, pp. 309–365. Paris (1903)
28. Reiter, R.: A theory of diagnosis from first principles. Artificial Intelligence **32**, 57–95 (1987)
29. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. In: IJCAI'17. pp. 2662–2670 (2017)
30. Shih, A., Darwiche, A., Choi, A.: Verifying binarized neural networks by Angluin-style learning. In: SAT'19. pp. 354–370 (2019)
31. Wang, Y., Zhang, Y., Zhou, Y., Zhang, M.: Knowledge forgetting in answer set programming. J. Artif. Intell. Res. **50**, 31–70 (2014)
32. Wang, Z., Wang, K., Topor, R.W., Zhang, X.: Tableau-based forgetting in [ascr][lscr][cscr] ontologies. In: ECAI'10. pp. 47–52 (2010)
33. Wernhard, C.: Literal projection for first-order logic. In: JELIA'08. pp. 389–402 (2008)
34. Wernhard, C.: Tableaux for projection computation and knowledge compilation. In: TABLEAUX'09. pp. 325–340 (2009)
35. Zhang, Y., Zhou, Y.: Knowledge forgetting: Properties and applications. Artif. Intell. **173**(16-17), 1525–1537 (2009)
36. Zhao, Y., Alghamdi, G., Schmidt, R.A., Feng, H., Stoilos, G., Juric, D., Khodadadi, M.: Tracking logical difference in large-scale ontologies: A forgetting-based approach. In: AAAI'19. pp. 3116–3124 (2019)
37. Zhao, Y., Schmidt, R.A.: FAME(Q): an automated tool for forgetting in description logics with qualified number restrictions. In: CADE'19. pp. 568–579 (2019)