From Explanations to Intelligible Explanations

Sylvie Coste-Marquis\textsuperscript{1} Pierre Marquis\textsuperscript{1,2}

\textsuperscript{1}CRIL, Univ Artois & CNRS
\textsuperscript{2} Institut Universitaire de France

XLoKR’20 Workshop, September 14\textsuperscript{th} 2020
Explaining is **hard** for a number of reasons

- It can be the case that an explanation is **useless** because it is **not intelligible**

- **Intelligibility** is not an **intrinsic property** of the explanation

- A **(user) model** associated with the explainee must be taken into account

- How to go from explanations to intelligible explanations?

- **KR** has developed some concepts (and tools) that can be useful to deal with such situations
With Abraham at the Ophthalmologist
Consider the following scenario:

- 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 (he suspects that he is myopic).
- 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)

The explanation is meaningful for the doctor, but not for Abraham

Among other things, Abraham would like to know whether it is hereditary
Explaining manifestations that are observed using a knowledge-based model

- $T$: a propositional formula (a domain theory)
- $A$: a subset of propositional symbols (the assumptions)
- $M$: a finite set of propositional formulae (the manifestations)
- $M^*$: a subset of $M$ (the manifestations to be explained)
Explaining manifestations that are observed using logical formulae

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

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

- **Minimal** explanations, i.e., explanations that are as weak as possible from a logical standpoint, are often considered

- Several preference criteria can be aggregated in order to define a notion of preferred explanation (e.g., minimality vs. coverage)
Explaining Abraham’s Manifestations

- $T = (ms \Rightarrow (bv \land ss \land he)) \land (my \Rightarrow (bv \land \neg he)) \land (ss \Leftrightarrow (ht \land lt))$
- $A = \{ms, my, co\}$
- $M^* = \{bv\}$ and $M = \{bv, ss\}$

- $ms$: "Abraham suffers from Marfan syndrome"
- $my$: "Abraham suffers from myopia"
- $bv$: "Abraham has a blurred vision"
- $ss$: "Abraham has the Steinberg’s sign (alias the thumb sign)” (a combination of hypermobility of the thumb ($ht$) as well as a thumb which is longer than usual ($lt$))
- $co$: "Abraham suffers from conjunctivitis"
- $he$: "Abraham suffers from a hereditary disease"
Explaining Abraham’s Manifestations

\\[ T = (ms \Rightarrow (bv \land ss \land he)) \land (my \Rightarrow (bv \land \neg he)) \land (ss \Leftrightarrow (ht \land lt)) \]

\\[ A = \{ms, my, co\} \]

\\[ M^* = \{bv\} \]

\\[ M = \{bv, ss\} \]

- \textit{ms, my} are minimal abductive explanations for \( M^* \) w.r.t. \( T \) and \( M \)
- The set of manifestations covered by \( ms \) is \{\( bv, ss \)\}
- The set of manifestations covered by \( my \) is \{\( bv \)\}
Projecting an Explanation onto a Vocabulary

- \( \gamma \): propositional formula (an explanation)
- \( U \): a subset of propositional symbols (the user vocabulary)
- \( T \): a propositional formula (a domain theory), that is supposed consistent

The projection of \( \gamma \) onto \( U \) given \( T \) is the set \( \Pi(\{\gamma\}, T, U) \) of all logical consequences over \( U \) of \( T \cup \{\gamma\} \)

- It is an infinite set

- Projection is not specific to the abductive model for explanations!
With Abraham at the Ophthalmologist (cont’ed)

- $\gamma = ms$
- Abraham would like to get all the information he may understand that are about this disease
- From the discussion she had with Abraham, the physician assumes that Abraham’s vocabulary contains $my$, $bv$, $he$ and concepts like $ht$ and $lt$ are common knowledge
- $U = \{my, bv, he, ht, lt\}$
- Then she may project $\gamma$ onto $U$ given $T$
- $\Pi(\{ms\}, T, U)$ is equivalent to $bv \land \neg my \land he \land ht \land lt$
- $bv$, $ht$, $lt$ can be filtered out assuming that Abraham knows those facts
- The physician can then explain Abraham that he does not suffer from myopia, and that unlike myopia, Marfan disease is hereditary
What about replacing $T$ by its projection onto the user vocabulary $U$ before computing explanations, or alternatively restricting $A$ to $A \cap U$?

This would not lead to the same set of explanations in the general case.

Hence the set of intelligible consequences that could be deduced from an explanation may heavily differ as well.
The projection of $T$ onto $U$ is equivalent to $my \Rightarrow (bv \land \neg he)$.

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 $M$.

Unlike $ms$, $my$ does not cover the manifestation $ss$ and for this reason, it has been considered as less preferred.

$my$ has consequences over $U$ given $my \Rightarrow (bv \land \neg he)$ that conflict with the consequences of $ms$ over $U$ given $T$.

$\neg he$ is a consequence of $my$ given $my \Rightarrow (bv \land \neg he)$ and $he$ is a consequence of $ms$ given $T$. 
- $\Pi(\{\gamma\}, T, U)$ is equivalent to the forgetting $\exists \bar{U}.(T \land \gamma)$ of $\bar{U}$ in $T \land \gamma$

- $\exists X.\phi$ is a quantified Boolean formula, equivalent to a formula that can be inductively defined as follows:
  - $\exists \emptyset.\phi \equiv \phi$
  - $\exists \{x\}.\phi \equiv \phi_{x \leftarrow 0} \lor \phi_{x \leftarrow 1}$
  - $\exists(\{x\} \cup X).\phi \equiv \exists X.(\exists \{x\}.\phi)$
Evaluating the Projection Operation: The Information Side

- Leads to an **information loss** in general:

\[ T \land \gamma \models \Pi(\{\gamma\}, T, U) \text{ but } T \land \gamma \not\models \Pi(\{\gamma\}, T, U) \]

- Projecting an explanation onto a user vocabulary can only **increase the amount of intelligible information** furnished to the user.

- Formally, suppose that the user also has her own knowledge base \( T_U \) (a propositional formula) such that \( U = \text{Var}(T_U) \), and \( T \models T_U \)

- We have

\[ \Pi(\{\gamma\}, T_U, U) \subseteq \Pi(\Pi(\{\gamma\}, T, U), T_U, U) = \Pi(\{\gamma\}, T, U) \]
Evaluating the Projection Operation:
The Explanation Side

- Needs to make precise the corresponding explanation model (here the abductive one)
- In the general case $\{T\} \cup \Pi(\{\gamma\}, T, U) \not\models M^*$

- $A = \{ms, my, co\}$, $M^* = \{bv\}$, $M = \{bv, ss\}$, and $U = \{my, bv, he, ht, lt\}$

- The physician prefers the explanation $ms$ to the explanation $my$ because it covers more symptoms than $my$

- The corresponding projection is equivalent to $bv \land \neg my \land he \land ht \land lt$ and the only conjunction of variables from $A \cap U$ that is consistent with it is the empty conjunction

- This assumption is consistent with $T$ but it does not explain $M^*$ (we have $T \not\models bv$)
It may happen that the explanations are not intelligible by the explainee but can be reformulated in terms of the explainee vocabulary.

This amounts to a definability issue.

An explanation $\gamma$ is definable in terms of the explainee vocabulary $U$ in the domain theory $T$ whenever there exists a formula $\Phi_U$ such that $\gamma$ is equivalent to $\Phi_U$ in $T$, i.e., we have $T \models \gamma \iff \Phi_U$.

When $\gamma$ is definable, any admissible $\Phi_U$ is referred to as a definition of $\gamma$ on $U$ in $T$. 
Abraham asks her doctor for a counterfactual explanation: why not considering myopia as an explanation?

The doctor then explains that Abraham also has the Steinberg’s sign, and myopia does not explain it.

Since $ss$ does not belong to $U$, once again, this explanation is not intelligible by Abraham.

However, $ss$ can be reformulated using Abraham’s vocabulary: $ss$ precisely means that Abraham’s thumb is hypermobile ($ht$) and longer than usual ($lt$).
When $\gamma$ is definable in terms of $U$ in $T$, one can project $\Phi_U$ onto $U$ given $T$ instead of projecting $\gamma$ onto $U$ given $T$:

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

This is helpful when the explainer knows that $T_U \models \exists \bar{U}. T$

Instead of providing $\Pi(\{\gamma\}, T, U)$ to the explainee, she can simply let her known as an explanation that $\Phi_U$ holds, and from it, the explainee will be able to deduce every piece of information conveyed by $\Pi(\{\gamma\}, T, U)$
In the abductive model for explanation, the projection of $\gamma$ onto $U$ given $T$ does not lead to an explainability loss when $\gamma$ is definable in terms of $U$ in $T$.

Suppose that $\gamma$ is an abductive explanation for $M^*$ w.r.t. $T$ and $M$ and that $\gamma$ is definable in terms of $U$ in $T$, so that there exists a formula $\Phi_U$ from $PROP_{PS}$ that is a definition of $\gamma$ on $U$ in $T$.

Let $\gamma_U$ be any implicant of $\Phi_U$ that is consistent with $T$. Then $\gamma_U$ is an abductive explanation for $M^*$ w.r.t. $T$ and $M$. 

Going from explanations to intelligible explanations requires to put in the explanation picture a model of the explainee.

A very basic user model, consisting of a logical vocabulary (a set of propositions which are meaningful), has been considered.

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.

The projection operation has been evaluated in terms of intelligibility, information, and explainability.

The specific case of definable explanations has been studied.

Theory reasoning can be used to simplify the explanations that are reported.
Next Steps

- Considering **more expressive settings** than classical propositional logic and investigating the extent to which the results presented in the paper can be lifted.
- Considering **other explanation models**.
- 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 an abundant literature (and some pieces of software).