# Requirements for Conceptual Representations of Explanations and How Reasoning Systems Can Serve Them

**Helmut Horacek**
*helmut.horacek@dfki.de*

## Abstract

Explanations of solutions produced by reasoning systems in ever growing complexity become increasingly interesting, which is particularly challenging in view of fundamental differences between human and machine representation and problem-solving methods. In this paper, we formulate requirements for conceptual representations that are adequate for producing human-oriented explanations, and we discuss how some reasoning mechanisms can serve them or can possibly be adapted to do so. This examination is intended to state in what ways reasoning systems can potentially support explanation generation, and where technology-justified limitations have to be accepted.

## 1 Introduction

Explanations justifying or questioning results produced by intelligent systems were always of interest, but this issue was rarely addressed in depth. Simple attempts with expert systems, although being one of the most explanation-friendly reasoning techniques, produced unexpectedly poor results, mainly because of insufficient understanding of impacts of human expectations and inference capabilities in discourse. This situation motivated the ambitious approach to *Explainable Expert Systems* (EES) (Swartout, Smoliar, 1997). The idea is to treat explanation not as an „afterthought“ but to foresee possible extra demands of explanations by incorporating suitable „built-ins“ within the reasoning process. While this strategy turned out to be successful for expert systems, it hardly looks promising for many other categories of reasoning systems where the discrepancy to human-like reasoning is considerably more pronounced.

An intellectual challenge has been mastered recently by a system that has beaten the human champion of Go in a match (Silver et al. 2016). The system applied deep neural network learning and searching on the basis of enormously large amount of data, that is, millions of games. While this is sufficient to outperform top level players in the purely performance-oriented task of a match, the system, similar to chess programs, cannot document its behavior in human-relevant terms, because it does not have an explicit representation of most domain-relevant concepts, which form the basis for human-adequate explanations.

Motivated by this gap between machine and human representation concepts, we formulate requirements for conceptual representations that are adequate for producing human-oriented explanations, and we discuss how some prominent reasoning mechanisms can serve them or can possibly be adapted to do so. Reasoning techniques referred to include rule-based representations, constraint-systems, decision trees, Bayesian networks and neural networks.

This paper is organized as follows. We first investigate what is required for producing explanations that are likely to be meaningful and useful to humans, and we formulate a set of complementary requirements. Then we examine some major reasoning techniques from the perspective of how they can serve explanation-motivated requirements, and if there is a gap, how it can possibly be narrowed. We also briefly address issues of natural-language presentation. Finally, we discuss the state-of affairs and expected future developments.

## 2 Requirements for Explanations

In this section, we discuss what is needed to provide representations from which human-adequate explanations can reasonably be generated with linguistic techniques. We focus on representations here because most linguistic presentation techniques needed to map these representations onto text are suitable for several genres of text. In addition, we devote a short section to specificities of explanations, such as the role of implicit conveyance of information, towards the end of the paper.



*Proceedings of the 1st Workshop on Explainable Computational Intelligence*, pages 11–15, Santiago de Compostela, Spain, September 4 2017. 

### 2.1 Categories for Explanations

Explanations may come in a variety of forms serving in part quite complementary purposes, mostly depending on the task at hand: this may be some proof of evidence for a solution, investigating constellations that may qualify for a solution, inquiring rationals for classification or decision preferences. We distinguish five categories of explanations:

1. *Exposition of the lines of reasoning*
   This kind of explanations addresses the adequate presentation of an inference chain, more general a tree or graph of inferences, be it in the context of a theorem prover, an expert system, or an argumentation framework. The purpose is to increase confidence in results obtained by a system, which may be by verifying the overall course of solution, or by referring to essential ingredients.
2. *Hypothetical inquiries*
   This kind of explanation is typically revelant for situations in which expectations or user beliefs are not met by solutions proposed by a system. Users may have interest in some specific constellation which turned out to be inferior or unacceptable, may be due to some small detail, or it may simply be unexplored. It is desirable for this kind of explanation to focus on essential reasons.
3. *Justification for categorization*
   This kind of explanation refers to the ingredients that have contributed to a taxonomic decision. As with the previous category, focusing on essential factors rather than on completeness is of importance here; thereby, possible reasons for misconceptons may be met, which may have motivated the explanatory request.
4. *Decision preferences*
   This kind of explanation refers to a comparison between properties of an entity in question and its closest competitors in the decision, or some explicitly mentioned candidates. In terms of what is to be compared, a combination of the previous and the following category may apply.
5. *Issues of calculation*
   This component of explanation focuses on impacts of quantitative properties and their dependencies on the issue to be explained. A detailed exposition of all calculations is of minor importance – simple ranges of numbers are preferable, including justification where they come from.

There are larger and richer catalogs of categories, but we think that we have captured the most principled ones. One important category missing are meta-explanations, about problem-solving strategies and their application, but this is a weakness of virtually all systems, since they do not have explicit representations of how they are working.

### 2.2 Properties of Human-Adequate Explanations

In order for representations to be suitable for explanations we think some criteria are indispensible:

1 *Focused content*
For an explanation to be useful, its content must be to the point of the purpose of the explanatory request. It is of little help if the content is somehow related to what is expected as an explanation. If the reasoning mechanism does not enable building an adequate response specification, we feel it is better to provide partial or incomplete information or evidence that may be not optimal.

2 *Vocabulary used*
The content provided should be expressed in terms the audience is familiar with. By this requirement, we do not mean a difference between expert and novice terminology, which can be bridged by natural language generation techniques, at least to some degree. The requirement addresses those cases where the problem-solving technique used by machines is fundamentally different from human approaches – human domain concepts are not used and may not easily be identifiable if at all within the machines' approach.

3 *Granularity*
The content of an explanation should be expressed in an adequate level of detail. If it is too detailed, an overall explanation may be longish and perceived as boring, and may even get incomprehensible. If it does not contain enough details, it may not be of use due to limited information.

## 3 Examining Explanation Potentials for Problem-Solving Techniques

In this section, we discuss to what extent the criteria elaborated in the previous subsection can be met by some major reasoning techniques, and we discuss measures to increase coverage and quality.

### 3.1 Systems with Rule-like Representations

This category comprises expert systems, automated theorem provers, and argumentation frameworks. These systems can serve the content of explanations for exposition of the lines of reasoning rather well. Similarly, the vocabulary and level of granularity is widely in accordance with human reasoning, except to automated theorem provers. They almost exclusively operate on the detailed resolution calculus, where the connection to the originally specified mathematical axioms, which constitute the basic vocabulary in this domain, gets widely lost. Fortunately, there are automated transformation procedures which can lift the proof representation to the more abstract *assertion level* (Huang, 1994).

An exception are cognitively involved inference patterns, such as modus tollens and disjunction elimination – several of these are composed into an assertion level step – a decomposition into natural deduction level steps is advisable (Horacek 1998, 1999, 2007). Some domain rules in expert systems may be associated with annotations that express their justification, much in the style of EES. Explanations on a higher level of granularity, such as as proof sketches and proof ideas are essentially unexplored. Skeletons of proofs plans may be adapted for this purpose, but such approaches are rare.

### 3.2 Constraint Systems

For this category of systems, the suitability for explanations appears to be rather good, at first sight. For typical applications, constraints themselves are expressed on a level corresponding to human views, so that vocabulary and granularity can be expected to be on a suitable level. May be, there are higher-level conceptions which correspond to a set or some composition of several constraints, so that addressing the more abstract view requires some transformation process to take place, possibly on demand by a specific explanatory request. Potential problems with explanations become only clearer when explanatory requests and information produced by the problem-solving techniques are put in relation to one another. Requests for justifying a solution are not of major interest for constraint systems; the simple explanation is just a message indicating that all constraints are fulfilled for the solution proposed. More informative messages would selectively list those constraints which are barely fulfilled, for constraints which involve a numerical comparison. In addition, a meta-explanation about the portion of the search space explored and the degree to which optimality is approached may be suitable, in case the system is set up in a way so that search is stopped when a solution with satisfactory quality is foumd. However, describing the search space in terms of which portions are still unexplored may be quite demanding.

Another category of explanations suitable for problems addressed by constraints systems are hypothetical inquiries. In a design problem, several inquiries may refer to partial constellations which the designer might expect or prefer to be part of a solution, but the system results show different combinations. In an explanation the reasons might be some violated set of constraints, but this information might not necessarily be complete or best. For excluding some combination of values from being a solution, a single constraint responsible for that is sufficient – in the search, every effort is made to exclude as much as possible on as little information as available. Hence, in order to obtain a more complete and focused view, checking and evaluating additional constraints for explanatory purposes only might prove suitable. An extreme approach for this purpose is described in (Horacek, 1992), which attempts to establishes dominances among sets of constraints, much in the style of Berliner (1979, 1982), but it requires full exploration of the search space. Altogether, explanations for constraint systems appear reasonably doable, but the content quality may not always be as desirable, and additional computation effort is required to address this issue.

### 3.3 Decision Trees

This problem-solving method is mostly suitable for obtaining categorizations or preferences between choices of some sort. Similar to constraint systems, (good) reasons for a possibly unexpected categorization, thus, a hypothetical solution are typical of interest. Conversely, major reasons for the categorization obtained are much more sensible here than a similar explanatory request in the conext of constraint systems. Structurally, the content of an explanation for a hypothetical solution is a description of the expressions of one of the choice points where the path to the category inquired is missed. Conversely, a complete description addressing a request for the categorization obtained comprises the expressions associated with all choice points on the path to that categorization. More focused explanations may choose a suitable one among the choice points in the first case and they may be selective in concentrating on conceptually more important ones in the second case.

In contrast to the previous two system categories, presenting the content of explanations may prove to be problematic here, since the expressions associated with the choice points may be quite complex, typically not corresponding to domain concepts meaningful to humans, since the overall tree structure is motivated by the goal of obtained a mostly balanced tree. Consequently, there is a serious problem in the vocabulary discrepancy between the components of decision trees and human domain conceptions. We are aware of only a single attempt to bridge this gap: in the domain of elementary chess pawn endings (king plus pawn versus king), decision trees were built to discriminate won from drawn positions (Michalski, Negri 1977). The tree learned on the basis of the board data only was compact, but its form was felt obscure by human players. When the building of choice point was biased by some force to use domain concepts, such as pawn square, king opposition, etc., the tree learned was structurally less optimal, but much better understandable to humans in terms of the discriminations made. This is a good example for explanations being a built-in, though in a different way as in EES.

| *Reasoning method* | *Weakness* | *Measures* |
|---|---|---|
| Rules | Granularity | Transformations |
| Constraints | Content, in part | Extra searching |
| Decision trees | Vocabulary | Biasing vocabulary |
| Bayesian networks | Role of numbers | Quantitative versions |
| Neural networks. | Content (+others) | Sensitivity analysis |

**Table 1.** Reasoning methods, their weak points in explanation, some measures against.

### 3.4 Bayesian Networks

In this category of systems, explanations may address generic or individual requests to the network. Generic requests concern the topology of the network, which comprises dependencies, justifications, and probabilities, possibly extended by annotations in the style of EES (e.g., giving sources or other details about the probabilities). Altogether, this is a presentation task pretty much on the lines of documenting rules, augmented by references to and descriptions of probabilities. Individual requests can be dealt with in more details. As far as the dependency of events is concerned, this amounts to a composition of rules, possibly in a tree. The extra component is the reference to and documentation of the probabilities associated with events and coocurrences of events. Merely listing the numerical data and the results of calculations is not difficult, but in some cases at least, there may be a better vocabulary in terms of qualitative assessments, as approximations. Such an approach has been undertaken in the context of argumentative presentations in natural language (Carenini, Moore, 2006), where the natural language descriptions were preferred by users to the precise graphical displays.

### 3.5 Neural Networks

This is clearly the most explanation-resistant technique described in this section. Its performance-oriented strength loses in explanation-related terms, since the important intermediate levels are not anywhere near a conceptual interpretation. Thus, the mathematical aspect is dominating, so that the architectural inspiration by the human brain somehow stops half way - the network learns how to perform, but does not produce explicit conceptions in the resulting representation. Consequently, there is virtually nothing that provides a basis for an explanation, only input and output data being on a level accessible to humans. Some more options are available for networks with a specific topology, such as gated networks (Zhao et al. 2017), where activations at intermediate levels can be visualized; but this techniques is probably suitable for a specific set of tasks only. What is remaining would be reruns with similar related data, to find out essential differences on some experimental basis. In addition, value differences between alternative output items could be used to refer to close competitors, e.g., near misses. However, how to orchestrate a reasonable set of recomputations effectively is ambitious.

A summmary of the reasoning techniques discussed, in terms of major weaknesses and meaures to potentially overcome them is given in Table 1.

## 4 Presentation Methods

Explanation presentation needs good sentence planning, including aggregation (Di Eugenio et al. 2005), and argumentation organization (Carenini, Moore, 2006). In addition, having a good command of explicitness and implicitness in presentation is of great importance in this genre (Horacek 1998, 2007), even more prominently in varions versions of the Digital Aristotle (Porter, 2007). Note that deliberately leaving portions of the content specification implicit is fundamentally different from selectivity in building content specifications: the latter means that they are not to be conveyed to the user, whereas the former is justified by the expectation that the audience is able to infer the content left implicit.

By and large, constellatons for leaving parts of content specification implicit are fairly well understood at a local level, such as the preference of modus brevis to fully exposed modus ponens presentations, straightforward taxonomic and action inferences, and expansion of known and mastered definitions. However, orchestrating the combination of several such constellations in a contextually adequate manner is still a widely unanswered question.

## 5 Conclusion and Discussion

In this paper, we have advocated in favor of necessary properties of representations that are suitable for specifics of explanations: the content, the vocabulary, and the level of granularity. We have discussed how these requirements are met or not met by some prominent reasoning mechanisms, We also have referred to measures already addressed and we have sketched some more ways to overcome existing discrepancies. Measures range from built-in methods to extra computations invoked by explanatory requests; they include transformation and enhancement of representations, and extra computations for parts not contributing to a solution.

Approaches to explanation require capabilities in several fields, such as automated theorem proving and NLP, which few researchers can cover. Nevertheless, increasing success and use of reasoning facilities will require a better documentation of their capabilities, especially for users who are sceptical towards machine-generated problem solutions.

## References


(Berliner 1979)
H. Berliner, On the Construction of Evaluation Functions for Large Domains. In Proceedings of the *6th International Joint Conference on Artificial Intelligence*, 53-55, Tokyo, Japan, 1979.

(Berliner, Ackley 1982)
H. Berliner, and D. Ackley. The QBKG System: Generating Explanations from a Non-Discrete Knowledge Representation. In Proceedings of the *Second National Conference on Artificial Intel- ligence (AAAI-82)*, 213-216, Pittsburgh, Pennsylvania, 1982.

(Silver et al. 2016)
D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel and D. Hassabis. Mastering the Game of Go with Deep Neural Networks and Tree Search. *NATURE*, Vol. 529, 2016.

(Carenini, Moore 2006)
G. Carenini and J. D. Moore, Generating and evaluating evaluative arguments. *Artificial Intelligence* Vol. 170, 925-952, 2006.

(Di Eugenio et al. 2005)
B. Di Eugenio, D. Fossati, D. Yu, S. Haller, and M. Glass, Natural Language Generation for Intelligent Tutoring Systems: a case study, *12th International Conference on Artificial Intelligence in Education*, Amsterdam, The Netherlands, 2005.

(Horacek 1992)
H. Horacek, Explanations for Constraint Systems. In B. Neumann (ed.), Proceedings of the *10th European Conference of Artificial Intelligence (ECAI-92)*, 500-504, Vienna, Austria, 1992.

(Horacek 1998)
H. Horacek. Generating Inference-Rich Discourse Through Revisions of RST Trees. In Proceedings of the *Fifteenth National Conference on Artificial Intelligence* (*AAAI-98)*, 814-820, 1998.

(Horacek 1999)
H. Horacek. Presenting Proofs in a Human-Oriented Way. In H. Ganzinger (ed.), Proceedings of the *16th International Conference on Automated Deduction* (*CADE-16)*, LNAI 1632, Springer, 142-156, 1999.

(Horacek 2007)
H. Horacek How to Build Explanations of Automated Proofs: A Methodology and Requirements on Domain Representations. *ExaCt 2007*: 34-41 Explanation-Aware Computing, Papers from the 2007 AAAI Workshop, Vancouver, British Columbia, Canada, July 22-23, 2007**.** AAAI Technical Report WS-07-06, AAAI Press 2007,

(Huang 1994)
X, Huang. Reconstructing Proofs at the Assertional Level. In Proceedings of the *12th International Conference on Automated Deduction (CADE-94)*, 738-752, Nancy, France, 1994.

(Michalski, Negri 1977)
R. S. Michalski and P. Negri An experiment on inductive learning in chess endgames.. In E. W. Elcock. D. Michie (eds.), Machine Intelligence 8, 175-192, Ellis Horwood, 1977.

(Porter 2007)
B. Porter. A New Class of Knowledge Systems and their Explanation Requirements. Invited talk at the *ExaCt 2007 Workshop on Explanation-aware Computing at AAAI 2007*, Vancouver, Canada, 2007.

(Swartout , Smoliar 1997)
B. Swartout and S. Smoliar. On Making Expert Systems more like Experts, Experts Sytsms, Vol 4(3), 196-208, 1997.

(Zhao et al.)
Y. Zhao, N. Semuma, X, Shen, and A.Aizawa. A Gated Neural Network for Sentence Compression Using Linguistic Knowledge, In Proceedings of 22nd *NLDB* conference, 2017.