Reconstructing Qualitative Model Variations from Qualitative Descriptions for Conversational Explanation

Moritz Bayerkuhnlein[a,b,*] **and Diedrich Wolter**[b]

[a]University of Bamberg, Germany
[b]University of Lübeck, Germany

Abstract. Qualitative reasoning models aim to capture how humans reason about common sense and real-world phenomena, yet not everyone has the same understanding, and thus underlying mental models of a phenomenon may differ. This paper presents a process for reconstructing qualitative models as proxies for capturing errors in a person's understanding. Using qualitative simulation models, we address situations where incorrect predictions are made, indicating gaps or errors in a person's understanding. Through an abductive reasoning process, we generate reconstructions of mental models that could reproduce these faulty predictions by adapting the expert model to reflect the person's perspective. Finally, we use the reconstructed models to formulate *contrastive explanations*, which aim to complete their mental model.

1 Introduction

In a conversation about a topic, participants rarely have exactly the same understanding of that topic. However, human communication is possible, even efficient, despite these differences in topic knowledge. This gap is most noticeable in a conversation between a teacher or expert and a learner.

The learner tries to puzzle out the relationships between the discussed concepts to build an understanding of the topic discussed. A good teacher will try to intuitively gauge the understanding of the student based on their (verbal) responses, to guide the conversation towards the desired learning outcome, and give relevant explanations. In other words, the teacher tries to understand the understanding of the student, asking the question: *How did they come to that conclusion?*

In this paper, we model this perspective taking using Qualitative Simulation Models as approximations of human mental models [11]. We assume an expert model on some given phenomenon, as well as a prediction made by a learner that is not compatible with the expert model, suggesting that the learner's *conception* is incomplete or misguided. We abduce potential models that explain the faulty prediction, adapting the expert model to a point where it captures the learner's lack of knowledge, or even *misconceptions* (see Figure 1).

Our approach is based on the foundations of Qualitative System Identification [28] and Abductive Diagnosis [7], yet does not construct a model from scratch, rather it builds adaptations from a reference model (the expert model mentioned above). If the learner's responses contain sufficient information, the resulting reconstructed model variant represents the deviation from the reference model to the faulty unknown model. Inferences made with the reconstructed model then provide insight into what information needs to be presented to correct the faulty mental model and inform the learner.

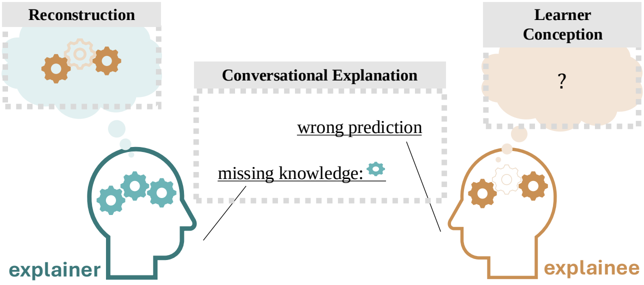


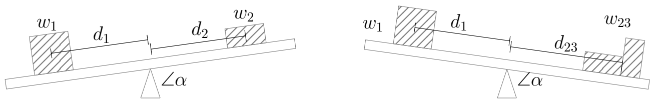
Figure 1. Overview: explainer (teacher) reconstructing the mental model of explainee (learner) before answering with relevant knowledge

Other approaches focus on models to learn qualitative behaviour from observations of systems, but here we are interested in articulate qualitative models that more closely resemble consistent human reasoning [12]. Reconstructing provides us with an interpretable model that can be used to assess the knowledge of the learner, generate hints, or, as will be discussed in Section 4, informs the generation of contrastive explanations [21, 15].

Furthermore, when considering a faulty physical system instead of a learner's *misconception*, the reconstructed model is a strong fault model for the device [7].

Running Example (Seesaw I)**.** *Consider the physical system of a seesaw. A student is asked to predict the behavior of the seesaw. He correctly states that it will tilt towards the heavier object w_1 (Figure 2a). The student is then told that an additional object w_3 of different weight is placed next to the lighter object w_1 such that the combined weight of w_2 and w_3 equals the weight of w_1 (Figure 2b). The student predicts that this will balance the seesaw, which is incorrect.*
Before providing an explanation, the teacher considers where the student's reasoning went astray, concluding it stems from either a lack of understanding of how the added object affects the center of mass or how it alters the lever's force.

[*] Corresponding Author. Email: moritz.bayerkuhnlein@uni-bamberg.de



(a) (b)

Figure 2. Seesaw configurations

2 Preliminaries & Related Work

The physical world can be described and reasoned about using precise mathematical equations and numerical information. However, humans tend to use qualitative information to reason and discuss phenomena of physical systems, reasoning and formulating qualitative arguments such as cause-effect relationships to convey the behavior of a phenomenon or system.

2.1 *Qualitative Simulation Models*

A perfect simulator would require complete knowledge of a situation and its dynamics. However, when dealing with verbal responses and mental models, we typically lack both. Instead, we must work with incomplete information and descriptions of multiple possible futures.

We build on Qualitative Differential Equations (QDE) as introduced by Kuipers [19]. QDE reduce a domain's quantitative constraints by representing only qualitative behavior, which is often more comprehensible and articulate than exact numerical representations.

Humans tend to base explanations on causal processes between physical entities, a model remains consistent with the domain's constraints but is more articulate by representing the cause-effect relationships between quantities in a comprehensible manner [12]. These relationships conceptually mirror human reasoning, reflecting how arguments about systems are phrased [8]. Explicitly modeling quantities and their qualitative causal relationships creates a *Qualitative Simulation Model* that can predict and explain system behavior using qualitative representations of differential equations and monotonic functions.

We follow the graphical realization of these models implemented in the Garp3 modeling toolkit [4]. The models are composed of *Ingredients*:

A qualitative simulation model is represented as a graph $\mathcal{QM} = \langle \mathcal{Q}, \mathbb{P} \rangle$, where:

- $\mathcal{Q}$ is a set of nodes representing *Quantities* associated with physical entities, with elements $q_1, \ldots, q_n \in \mathcal{Q}$.
- $\mathbb{P}$ is a set of (directed) edges representing *Processes*, which indicate causal dependencies between quantities.

Additionally, qualitative models can incorporate observations OBS, here we consider observed values of quantities or the relations between them.

Quantities Q can occupy a range of values expressed through a range of coarse mappings to a domain $\mathbb{D}(q), q \in Q$ called *quantity spaces*. At any given discrete time point t_i where $1 \leq i \leq h \in \mathbb{N}$, each quantity q has a value $\text{val}(q, t_i) \in \mathbb{D}(q)$ and a derivative $\delta(q, t_i) \in \{-, 0, +\}$. The derivative indicates the trend of the quantity at the next time point t_{i+1}.

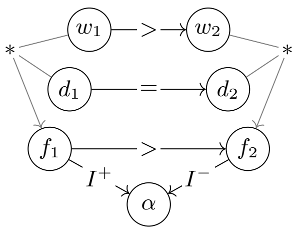
Processes $\mathbb{P}$ are labeled edges between two quantities q_i, q_k, taking the role of causal dependencies and determining the result of a simulation by constraining and influencing the values of the quantities. Between a quantity q_i and a target quantity q_j, causal dependencies take the form of **Influences** $I^{\pm}(q_i,\ q_j)$, which cause the target quantity q_j to change its derivation based on the magnitude of q_i, **Proportionalities** $P^{\pm}(q_i, q_j)$ operating as indirect influences propagating the effect of a process from q_i, to q_j, and **Correspondences** $Q(q_i, q_j)$, where the magnitudes of quantities correspond. In addition, a quantity can act as an auxiliary variable and be related to values calculated from other quantities using a **Calculation** here limited to multiplication and subtraction denoted by operations $q_i * q_j = q_k$ and $q_i - q_j = q_k$, respectively.

The dynamics of the simulation are determined by influences, proportionality, correspondences, etc., where causal dependencies determine the derivative $\delta(q, t_i)$ and the value $val(q, t_i)$ of each quantity. The collection of all derivatives and values at a given time point is called **State** s. A sequence of states modeled by the qualitative simulation model is called **Scenario** π.

Observations are concrete values obtained, for example, by measuring quantities or through a verbal description of a scenario we wish to simulate. A qualitative simulation can be constrained by **Assumptions** made about the configuration of the system. **Inequalities** $\{>, =, <\}$ between quantities are used to enforce constraints in the form of a relative position on a quantityspace, they can be enforced as constraints, or their truth values can act as additional observations to a scenario. A model constrained by an assumption must realize it at a specified time during the simulation (postdiction).

Finally, given an initial state s_0, a qualitative simulation model yields a **State Graph** Π consisting of states and transitions between these states. By traversing the graph, every possible simulation outcome (scenario) can be obtained. Thus, given a set of assumptions, there is a state *sub-graph*, which only includes scenarios consistent with the observations. If there is not a single state within the subgraph, then the qualitative simulation admits to no consistent scenario, and we speak of a contradiction.

Running Example (Seesaw II)**.** *Consider the seesaw in Figure 2a, with a central pivot point and two loaded arms with weights. The angle α of the seesaw, as well as the load w_1, w_2 and position d_1, d_2 of the weights are represented as quantities and relations and can be observed, e.g., $w_1 > w_2$. Finally, the lever force is not directly observable, but can be determined, represented here by f_1, f_2 pressing down on the respective sides. Figure 3 shows a graphical representation of an expert $\mathcal{QM}$ which realizes the dynamics of the seesaw, by considering the lever effect with f_1, f_2, which influence $I^+(f_1, \alpha), I^-(f_2, \alpha)$ the angle of the seesaw, as edges between the quantities.*



(a)

$\mathcal{Q}$	**quantityspaces**
d	$\{0, +, \max\}$
w	$\{0, +, \max\}$
f	$\{\min, -, 0, +, \max\}$
α	$\{\min, -, 0, +, \max\}$
$\mathcal{Q} \times \mathcal{Q}$	$\{<, =, >\}$

(b)

Figure 3. Qualitative Model of a Seesaw depicted in Figure 2a

In the scenarios of this models simulation, the quantities w, d, and f remain unchanged with a positive value each. Starting from a balanced seesaw in the initial state s_0, the scenario π is generated such that $val(\alpha_0, t_0) = 0, \delta(\alpha_0, t_0) = 0 \subset s_0 \mid s_0 \in \pi$. The force f_1 exerts a stronger influence than f_2, because the weight w_1 of the left object is heavier and f is calculated by $w * d = f$. This causes a *transition in the seesaw's state to $\delta(\alpha, t_1) = +$. This results in an update of $val(\alpha, t_1) = 0$ to $val(\alpha, t_2) = +$, indicating that when t_2 is reached, the seesaw is tilted to the left, with a positive angle. Eventually, the magnitude of the quantity α will converge to the maximum value $val(\alpha, t_n) = max$.*

2.2 Qualitative Model Abduction

A *System Identification Problem* is the process of using observations to understand the underlying structure of a system. This can be used to post hoc interpret the way a system works by reconstructing it as a model [1].

We speak of *Qualitative System Identification*, when we use qualitative modeling and observations to abduce a model that explains the behavior of the observed systems automatically [28].

For *Qualitative Differential Equations* (QDE), which model the dynamics of a system as a conjunction of qualitative constraints, the term *QDE model learning* (QML) refers to the inverse of qualitative simulation. Instead of predicting an outcome, in QML a model is induced from observation [23]. QDE model learning has been used to learn form observations of a physical target system. In qualitative reasoning, most automatic model construction approaches try to generate models that describe the behavior of the system using qualitative differential equations [29]. In general, they follow an abductive principle of hypothesis generation and pruning of inconsistent models. The approaches include GENMODEL [6], QSI [28] and MISQ [26]. Others rely on Inductive Logic Programming (ILP) [22] as a framework for model synthesis, also benefiting from the available systems to learn from both positive and negative examples [3, 5].

Abduction is the inference to the best explanation. While QED capture the qualitative dynamics of a system, they do not have the same articulate power as an explicit representation of processes and causal dependencies [12]. When reconstructing models from observation to understand erroneous behavior, we speak of *abducing* a qualitative model [17]. For the qualitative simulation models $\mathcal{QM}$ repesented by graphs, we can formally specify the problem as a general inductive problem [22]:

Definition 1 (Qualitative Model Abduction Problem)**.** *Given observations* OBS *and the dynamics of qualitative simulation S, reconstruct a model $\mathcal{QM} \subseteq \mathcal{L}_{\mathcal{QM}}$ by induction from a language of possible ingredients $\mathcal{L}_{\mathcal{QM}}$. The goal is to find $\mathcal{QM}$ such that:*

$$S \cup \mathcal{QM} \vdash \text{OBS} \tag{1}$$

In other words, we abduce a qualitative model $\mathcal{QM}$ that in accordance with the governing simulation rules S can reproduce the observations OBS that arise from a dynamic system under observation. The constructed model thus is said to *justify* the observations.

Precise parameterized models are hard to learn because of infinite possibilities in parameter assignments. Qualitative models abstract from these mathematical details, yielding only finite possibilities. They are easier to learn and can capture the dynamics of the system while remaining comprehensible. However, naive construction can lead to under- or over-constrained models, potentially causing faulty predictions [5].

2.3 Diagnosis

Conceptually reconstructing a model for a system that deviates from expected behavior can be framed as a *Diagnosis Problem* [25], where we search for a diagnosis Δ as a set of *abnormal* components to explain and ultimately repair faults within the system.

Definition 2 (Diagnosis Problem Instance)**.** A diagnosis problem instance consists of a triple, $\langle \text{SD}, \text{OBS}, \text{COMP} \rangle$

- system description (SD) , specifying the behavior and structure;
- a set of observations (OBS) on the system as facts;
- a set of constants c_i, representing the components (COMP).

The dominant approach to Model-Based Diagnosis is called Consistency-Based Diagnosis and has been successfully applied to Qualitative Simulation Models in [8]. Consistency-Based Diagnosis characterizes the behavior of a faulty component using only a binary label to indicate whether a component is *abnormal* or *ok*, forming sets of abnormal components, the diagnoses Δ [25].

When so-called *strong fault models* are available, the abductive approach to diagnosis can be used [7, 24]. Here the behavior of the faulty components is modeled in the diagnosis Δ and justifies the observations, such that

$$\begin{aligned} &\text{SD} \cup \Delta \vdash \text{OBS}, \\ &\text{SD} \cup \Delta \text{ is consistent} \end{aligned} \tag{2}$$

These fault models are however not easily obtained, as they generally rely on expert knowledge or existence of a bug-catalog. If a system description guarantees that even abnormal components operate on values confined to a specified domain (such as a quanitity space) constraints can be enforced. These constraints can be used to infer potential input-output behaviors even in the absence of an explicit strong fault model [2].

In a simulation, these reconstructed input-output values are placed between each state transition but are fundamentally governed by the dynamics of the system model. Finally, we want to point out that reconstructing the simulation model as the generator of these states can potentially also be revealing for diagnostic purposes.

3 Reconstructing Faulty Simulation Models

A qualitative simulation model is faulty if it cannot reproduce the behavior of an observed phenomenon. When representing something as illusive as the mental model of a learner, this qualitative simulation is rather abstract and hidden. From now on, we refer to this *abnormal* and hidden model as the *learner* model $\widetilde{\mathcal{QM}}$.

Presumably, for any observed phenomenon, there is a perfect model which captures exactly the dynamics required; we will refer to this correct model as the *reference* model $\mathcal{QM}$.

In our method, we start from an informed model and regress it by inducing model ingredients which explain a prediction made by an uninformed model. The resulting model is a reconstruction $\widehat{\mathcal{QM}}$, which acts as an approximation of the uninformed model.

More formally, we perform an abductive diagnosis, by reconstructing the model form a language of ingredients $\mathcal{L}_{\mathcal{QM}}$ such that:

$$S \cup H \cup (\mathcal{QM} \setminus R) \vdash \text{OBS}, \tag{3}$$

$$S \cup H \cup (\mathcal{QM} \setminus R) \text{ is consistent} \tag{4}$$

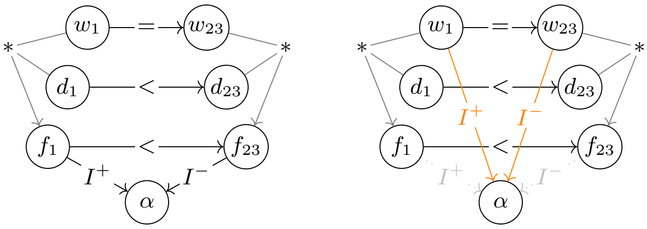
$$|H \ \cup R\,| \text{ is minimal} \tag{5}$$

where $R \subseteq \mathcal{QM}$, $H \subseteq \mathcal{L}_{\mathcal{QM}}$ and $\widehat{\mathcal{QM}} = H \cup \mathcal{QM} \setminus R$. Reconstructed models $\widehat{\mathcal{QM}}$ are instances of the language power set $\mathcal{L}_{\mathcal{QM}}$, $\mathcal{P}(\mathcal{L}_{\mathcal{QM}})$. The parsimony principle modeled in Equation 5 favors reconstructions to be close to the reference model $\mathcal{QM}$.

Intuitively, we adapt the reference model by retracting (R) and hypothesizing (H) model ingredients to account for the observations. In this context, observations OBS are not derived from measurements of the physical world. Instead, they are the products of predictions made by $\widetilde{\mathcal{QM}}$ (by the learner), which provide partial descriptions of states. These observations are presented as values or truth values of relationships of quantities.

The problem of constructing a consistent model from an empty reference Model $\mathcal{QM}$ where $\mathbb{P} = \emptyset$ is identical to the qualitative model abduction with Equation 1.

Running Example (Seesaw III)**.** *Consider the seesaw from Figure 2b. We can reuse the reference model from Figure 3a substituting with w_{23}, d_{23} and f_{23}, the configuration is depicted in the Figure 4 below. This presumes that the learner did not make a mistake interpreting the scene.*



(a) (b)

Figure 4. Qualitative Model of a Seesaw depicted in Figure 2a

A learner's prediction like "the seesaw will be balanced" can then be stated as an observation on their hidden learner model $\widetilde{\mathcal{QM}}$. An observation is a partial state description $\{val(\alpha, t_n) = 0\}$. This is underconstrained, a more restrictive interpretation of the utterance is $\forall 0 \leq h \leq n \colon \neg(val(\alpha, t_h) \neq 0)$, denoting the constraint that no state in this scenario may ever have an unbalanced seesaw. A model that reconstructs this, while also realizing the scenario without the additional object, is depicted in Figure 4b. Note that α is not influenced by the forces, but directly by the weights.

3.1 *Model Adaptation Language*

The reconstructed model is derived using a transformation language $\mathcal{L}_{\mathcal{QM}}$. During the adaptation, we do not consider adding auxiliary quantities to the reconstructed model. Instead, the adaptation language $\mathcal{L}_{\mathcal{QM}}$ consists of the processes $\mathbb{P}$ described in Section 2.1. This ensures that the potential adaptations expressible with $\mathcal{L}_{\mathcal{QM}}$ are finite.

Formally, the language is constructed from graph edit operations on the model ingredients performed on $\mathcal{QM}$. Here, these operations are limited to *edge insertion*, hypothesize a process between quantities, and *edge deletion*, retracting a process from the reference model.

Listing 1. Extended Backus–Naur form (EBNF) for Adaptations

```
<q> ::= q ∈ 𝒬

<adaptation> ::= <edit> " " <adaptation>
               | <edit>

<edit> ::= "delete" "(" <ingredient> ")"
         | "insert" "(" <ingredient> ")"

<ingredient> ::= <process>
               | <correspondence>
               | <calculation>

<process> ::= "I" <sign> "(" <q> "," <q> ")"
            | "P" <sign> "(" <q> "," <q> ")"

<correspondence> ::= "C(" <q> "," <q> ")"
                   | "C⁻¹(" <q> "," <q>")"

<calculation> ::= <q> "*" <q> "=" <q>
                | <q> "-" <q> "=" <q>

<sign> ::= "+" | "-"
```

Listing 1 presents a grammar for generating sequences of graph edits, representing sets H for insertions and R for deletions. A person who incorrectly assumed some causal dependency might have what is referred to as a *misconception*, which here is represented as the set H. However, failure to apply some knowledge is modeled as R.

If more involved edit operations are used, the minimality constraint on the model adaptations in Equation 5, can be revised using *graph edit distance GED* [27] between the reference and the reconstructed model such that:

$$\min_{qm \in \mathcal{P}(\mathcal{L}_{\mathcal{QM}})} GED(\mathcal{QM}, qm) \tag{6}$$

We are motivated to abduce models that minimize the edit distance to a reference model during reconstruction, since *conceptions* of learners in a learning situation is usually guided, also in the context of $\mathcal{QM}$ [18]. *Misconceptions* that deviate stronger from the intended reference model are possible, especially when learners rely on their intuition from past experiences and expertise in other domains [30, 8]. As such, the edited distance proposed here acts as one of many potential heuristics to find a good reconstructed model. For example, another heuristic might be informed based on the analogical reasoning and related knowledge the learner might possess [13].

4 Conversational Explanation

Abduction as the inference to the best explanation of an explanandum is only part of the explanation process. An explanation is fundamentally contextual, as it serves as a response to a question within a specific context [31]. In the conversation between the explainer and the explainee, this context is largely the *epistemic* state of the parties.

There are many aspects that factor into how humans converse, such as *quality*, *quantity*, and *manner* [14].

There are many aspects of how people converse that are summarized in Grice's Maxims of Conversation, such as ensuring that what is said is true (*quality*), that what is said is only as informative as necessary (*quantity*), and that statements are clear and understandable to the receiver (*manner*) [14]. Here we want to focus on the *relevance* of the logical content of a possible explanation in order to expose information for the repair of the explainees *epistemic* state.

4.1 *Contrastive Explanation*

Explanations in conversation are formulated against *Why*-questions. However, explainers will refrain from exposing unnecessary information and instead formulate an answer against an implied counterfactual alternative, which can also be made explicit by explainee as a "Why *explanandum* (ϕ) rather than *foil* (ψ)"-question. A response to such a question is called *contrastive explanation* [16].

A faulty prediction by a learner establishes a natural contrast to the informed prediction. A prediction from a learner that states ψ, acts as a counterfactual that stands in contrast to the actual true answer ϕ. Furthermore, since the learner had to generate the utterance from an *epistemic* state, the foil ψ also acts as an observation OBS and a basis for abduction of said *epistemic* state.

4.2 Explanations from Qualitative Simulation Models

The generation of intuitive explanations is one of the main concerns of qualitative models [10]. Since causal dependencies are modeled explicitly and are fundamental to the simulations dynamics, a simulator can also track the inferences made to reach a state, leading to a causal chain. Without special points of focus on these chains, the explanations naively will retrace the inference from initial state to the explanandum. Here we want to adapt the computational models of *contrastive explanation* from causal models [21] and logic programs [9], to fit Qualitative Simulation Models.

Definition 3 (Explanation Frame)**.** *An Explanation Frame* $\mathcal{F} = \langle \mathcal{QM}, s_0, S, \mathcal{L}_{\mathcal{QM}} \rangle$ *where*

- $\mathcal{QM}$ *is a reference model,*
- s_0 *a (partial) starting state,*
- S *the set of shared knowledge, and*
- $\mathcal{L}_{\mathcal{QM}}$ *the language for the hypothesis space.*

Definition 4 (Contrastive Explanation Problem)**.** *Given an explanation frame* $\mathcal{F} = \langle \mathcal{QM}, s_0, S, \mathcal{L}_{\mathcal{QM}} \rangle$*, a corresponding* Contrastive Explanation Problem *is a* $\mathcal{P} = \langle \pi, \phi, \psi \rangle$ *where*

- π *is a scenario of* $\mathcal{QM}$ *representing the actual prediction of* $\mathcal{QM}$*,*
- $\phi \subseteq \pi$ *is the* explanandum*, and*
- ψ *represents the* foil *with* $\psi \cap \pi = \emptyset$*.*

We use the foil ψ as an observation OBS for the reconstruction of $\widetilde{\mathcal{QM}}$, we obtain a reconstructed model $\widehat{\mathcal{QM}}$ as outlined in Section 3, as well as the divergence form the reference model as $H \cup R = Q_\Delta$, in practice multiple responses could be considered to improve the reconstruction. We collect the sets of causal dependencies and causal inferences Q_ϕ and Q_ψ that contributed to $S \cup \mathcal{QM} \vdash \phi$ and $S \cup \widehat{\mathcal{QM}} \vdash \psi$ respectively. Both Q_ϕ and Q_ψ are causal-explanations, using inference rules of the Qualitative Simulation, composed out of the state transitions with reference to the used model ingredient. For example, a simulation rule of S using a *Influence*-Ingredient.

The rule given in R1 below shows how the presence of a positive influence I^+ between two quantities q_1 and q_2 possibly changes the derivation of q_2 from one to the next time point. Conceptually, the model ingredients act as toggles of specific instantiations of rules within the logic program.

$$\begin{aligned} &\delta(q_2, i, +) \leftarrow \\ &\qquad I^+(q_1, q_2), \\ &\qquad \delta(q_1, i-1, \delta_{i-1}), val(q_1, i-1, v), v > 0. \end{aligned} \tag{R1}$$

If R1 is used during the simulation we record the model ingredient as a justification, indexed by the time point of use in Q_ϕ, Q_ϕ respectively.

Running Example (Seesaw VI)**.** *A simulation spanning timepoints* t_0*,* t_1 *and* t_2 *starting with* $\{val(\alpha, t_0) = 0, w_1 > w_2\} \subset s_0$ *with weight placed as depicted in Figure 2a on a neutral seesaw, realizing the* explanandum $\phi = \{val(\alpha, t_2) = +\}$ *cites* $Q_\phi = \{I^+(f_1, \alpha)_{t_0}\}$ *as an explanans, as an application of Rule R1.*

Finally a *contrastive explanation* can be obtained by contrasting both of the explanations Q_ϕ and Q_ψ as defined in [9].

Definition 5 (Contrastive Explanation)**.** *A counterfactual explanation* $\langle Q_\phi, Q_\psi, Q_\Delta \rangle$ *for an explanation frame* $\mathcal{F}$ *is made contrastive* $\langle C_\phi, C_\psi, C_\Delta \rangle$ *only when considering deviations and excluding shared knowledge* S*.*

- $C_\phi = Q_\phi \setminus (Q_\psi \cup S)$
- $C_\psi = Q_\psi \setminus (Q_\phi \cup S)$
- $C_\Delta = Q_\Delta \setminus S$

The parts of the contrastive explanation $\langle C_\phi, C_\psi, C_\Delta \rangle$ here denote the root-cause C_Δ of the faulty inference made by the explainee, and the resulting divergence in their reasoning C_ψ. The explanation carrying the information for a repair of the explainee's understanding is C_ϕ, outlining the explanation of the reference model $\mathcal{QM}$, reduced to the relevant inferences $\widehat{\mathcal{QM}}$ could not make due to the divergence.

A conversational verbalization of the contrastive explanation could, for example, cite the root cause C_Δ and give the retracing of actual inferences of C_ϕ.

5 Experiment

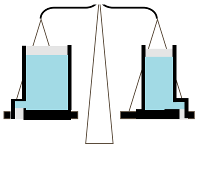
We have implemented qualitative simulations using graph models in Answer Set Programming as a prototype, where the dynamics of the simulations is encoded in rules such as R1 in a logic program. The implementation can generate scenarios, complete partial states to complete states, and generate a full state graph using *brave* enumeration, realizing *prediction*, *postidiction* and *causal reasoning* [12].

To illustrate the results of this approach, we will give a example used in education, where a faulty prediction will prompt reconstruction and explanation of the discrepancy.

Although dedicated ILP tools are available for learning answer set programs such as ILASP [20], they do not scale to the search space required for the full reconstruction of $\widehat{\mathcal{QM}}$ yet. For this example, we limit the adaption language $\mathcal{L}_{\mathcal{QM}}$ to only consider edge deletions.

5.1 Balance Domain

The following example shows a revised version of deKonning and Bredeweg's balance system [8] implemented as a graph model. The original version applied model-based diagnosis to diagnose the reasoning steps taken by a learner to generate feedback. With the use of a reconstructed articulate model and inherent explanation, we want to build on that.



(a)

Variable	Value
$L_L \times L_R$	$<$
$V_L \times V_R$	$=$
$W_L \times W_R$	$>$
pos	$neutral$

(b)

Figure 5. Balance with filled containers in initial configuration with water level (L), volume (V) and width (W). And auxiliary values that are not apparent from the static image, such as flow (F). Relative and qualitative values are made explicit here, including the angle of balance $angle \in \{left, neutral, right\}$

Consider the *sketch* shown in Figure 5 of a balance scale with two full containers. From the picture, we can make qualitative observations, such as comparing the water levels of the containers (c), or

Table 1. Partial valuations of scenarios of the quantities and relations from qualitative simulation of Figure 5. An actual scenario (left) according to the qualitative model. A counterfactual scenario (right) that accounts for an observation $V_L = V_R$ for all timepoints t

	$\mathcal{QM}$-Scenario				$\widehat{\mathcal{QM}}$-Scenario			
	t_0	t_1	t_2	t_3	t_0	t_1	t_2	t_3
V_L	$1, \rightarrow$	$1, \downarrow$	$0, \rightarrow$	$0, \rightarrow$	$1, \rightarrow$	$1, \downarrow$	$0, \rightarrow$	$0, \rightarrow$
V_R	$1, \rightarrow$	$1, \downarrow$	$1, \downarrow$	$0, \rightarrow$	$1, \rightarrow$	$1, \downarrow$	$0, \rightarrow$	$0, \rightarrow$
$V_L \times V_R$	=	=	>	=	=	=	=	=
$L_L \times L_R$	<	<	<	=	<	<	=	=
$F_L \times F_R$	<	<	<	=	=	=	=	=
$P_L \times P_R$	<	<	<	=	=	=	=	=
$W_L \times W_R$	>	>	>	>	>	>	>	>

qualitatively determining whether the scale (b) is tilting left or right. Opening the valves sets in motion a chain of events: the mass (m) of the containers, which depends on the volume (V), which depends on the water level (L), which regulates the pressure (p), which regulates the outflow (f), which influences the volume, which influences the outcome of the scales (pos).

As an example, we formulate an utterance from a student recorded in [8]. The student had been asked about the situation in Figure 5 where the containers start with the same volume: "Both valves are opened simultaneously. How will the volumes behave?".

The right-hand side, will have faster outflow, but a *wrong* prediction that does not consider the pressure within the containers could be: "The volumes of the remaining water will decrease equally, staying in the same relation." The answer suggests an observation $V_L = V_R$ for all time points $t_1, \ldots, t_n$ and both δV_L and δV_R are negative. This cannot be achieved by any scenario within the state graph of $\mathcal{QM}$. Adaptations are searched to find a reconstruction $\widehat{\mathcal{QM}}$.

By contrast, the reference model $\mathcal{QM}$ can predict the actual outcome, "The volume of the right containers will empty faster".
Framing this exchange as a Why-Rather-Than-Question, we get: "Why will the volume of the right containers decrease more quickly, rather than both decreasing equally?".

Among the minimal sets of deletion edits made to $\mathcal{QM}$ to generate $\widehat{\mathcal{QM}}$ which models the student's utterance are $R_1 = \{C(p, f)\}, R_2 = \{C(l, p)\}$, both adaptations can lead to a scenario outlined in Table 1. Either $C(l, p)$, the student has not considered the correspondence between the water level (l) and the pressure (p), or $C(p, f)$, they have not considered the correspondence of pressure (p) on flow out (f). The contrastive explanation obtained from the model where $C(l, p)$ is retracted is as follows:

$$C_\phi = \{C(l,p), C(p,f), C(l_R, p_R)_{t_2}, C(p_R, f_R)_{t_2}\}$$
$$C_\psi = \{C(v_L, m_L)_{t_2}, I^-(m_L, pos)_{t_1}, I^+(m_R, pos)_{t_1}\}$$
$$C_\Delta = \{C(l,p)\}$$

The indexed items, reference states within the scenario that the simulation generated (see Table 1). Interpreting the logical content of the explanation could yield the following, starting with the root-cause: *The right container's volume decreases quicker, because the water pressure corresponds to the water level ($C(l, p)$). At some point (t_2), the outflow from the right container is larger than from the left container ($F_L < F_R$), because the right container has a higher water level ($L_L < L_R$), and pressure and outflow are proportional ($C(p, f)$).*

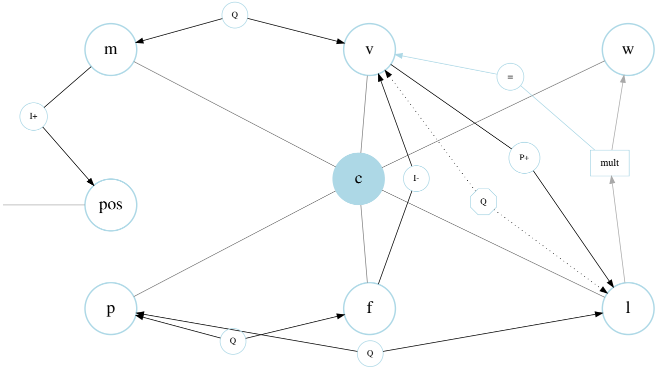


Figure 6. Qualitative Simulation Model excerpt of right container (c) as reference model $\mathcal{QM}$. Q-nodes denoting correspondences, dotted lines conditionals and white node quantities

5.2 *Limitations & Future Work*

Currently, our system does not realize learning from negative examples efficiently. Unlike the observation of a physical system, where only positive examples are produced, a human utterance can, in fact, carry information about a negative example, or be implied, as we have shown in the running example. Comparable general systems such as ILASP implement learning from negative examples using *cautious* consequences, but these systems are not scalable to the task of reconstructing a qualitative simulation model in a graph representation, as we have learned.

The constraint in 4 regarding inconsistencies of reconstructed models might not be realistic when it comes to human reasoning, as human reasoning often uses heuristics or accepts inconsistencies in order to act faster. An appropriate suspension of this constraint must be investigated in the future.

To handle the reconstruction of larger models, future work will invest in a dedicated method for abduction models, benefiting from advances in the field of constraint and inductive logic programming.

6 Summary & Conclusion

Explanation is the process of resolving a puzzle in the explainee's mind by filling gaps in their knowledge. However, each individual's mind is unique and not directly observable. Nevertheless, much like observing a system, the questions and answers provided by the explainee can serve as indicators of their flawed mental model.

In this work, we tackled the challenge of reconstructing qualitative model variations from responses to provide effective conversational explanations. Qualitative Simulation Models have been emphasized as a useful tool for addressing inconsistencies in predictions and capturing the way humans articulate their reasoning about processes. Using abductive and inductive reasoning, we can construct qualitative models from faulty predictions. This involves reconstructing mental models that adapt expert models to reflect the learner's perspective. This approach aims to bridge the understanding gap between teachers or experts and learners, ultimately improving learning outcomes and facilitating more effective explanations.

Additionally, the use of contrastive explanations formulated with the reconstructed models helps to complete the understanding of a person's mental model. By framing explanations in terms of Why-Rather-Than-Questions, we can gain insights into the reasoning behind different perspectives. This method provides a deeper understanding of the explainee's thought processes and helps tailor explanations to address specific misunderstandings.

Acknowledgements

The work presented in this paper has been carried out in context of the VoLL-KI project (grant 16DHKBI091), funded by Bundesministeriums für Bildung und Forschung (BMBF)

References

[1] K. J. Åström and P. Eykhoff. System identification—a survey. *Automatica*, 7(2):123–162, 1971.

[2] M. Bayerkuhnlein and D. Wolter. Model-based diagnosis with asp for non-groundable domains. In *International Symposium on Foundations of Information and Knowledge Systems*, pages 363–380. Springer, 2024.

[3] I. Bratko, S. Muggleton, and A. Varšek. Learning qualitative models of dynamic systems. In *Machine Learning Proceedings 1991*, pages 385–388. Elsevier, 1991.

[4] B. Bredeweg, F. Linnebank, A. Bouwer, and J. Liem. Garp3—workbench for qualitative modelling and simulation. *Ecological informatics*, 4(5-6):263–281, 2009.

[5] G. M. Coghill, A. Srinivasan, and R. D. King. Qualitative system identification from imperfect data. *Journal of Artificial Intelligence Research*, 32:825–877, 2008.

[6] E. W. Coiera. *Generating qualitative models from example behaviours*. Department of Computer Science, School of Electrical Engineering and ..., 1989.

[7] L. Console and P. Torasso. A spectrum of logical definitions of model-based diagnosis 1. *Computational intelligence*, 7(3):133–141, 1991.

[8] K. De Koning, B. Bredeweg, J. Breuker, and B. Wielinga. Model-based reasoning about learner behaviour. *Artificial Intelligence*, 117(2):173–229, 2000.

[9] T. Eiter, T. Geibinger, N. H. Ruiz, and J. Oetsch. A logic-based approach to contrastive explainability for neurosymbolic visual question answering. In *Proceedings of the 32rd International Joint Conference on Artificial Intelligence (IJCAI 2023)*, 2023.

[10] K. Forbus and B. Falkenhainer. Self-explanatory simulations: An integration of qualitative and quantitative knowledge. *Faltings & Struss*, pages 49–66, 1992.

[11] K. Forbus and D. Gentner. Qualitative mental models: Simulations or memories. In *Proceedings of the eleventh international workshop on qualitative reasoning*, pages 3–6. Citeseer, 1997.

[12] K. D. Forbus. Qualitative process theory. *Artificial intelligence*, 24 (1-3):85–168, 1984.

[13] S. Friedman and K. D. Forbus. Learning naïve physics models and misconceptions. In *Proceedings of the 31st annual conference of the cognitive science society*, pages 2505–2510, 2009.

[14] H. P. Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

[15] R. Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.

[16] D. J. Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65, 1990.

[17] I. C. Kraan, B. L. Richards, and B. Kuipers. Automatic abduction of qualitative models. In *Proceedings of the Fifth International Workshop on Qualitative Reasoning about Physical Systems*, volume 295, page 301, 1991.

[18] M. Kragten, T. Hoogma, and B. Bredeweg. Learning domain knowledge and systems thinking using qualitative representations in upper secondary and higher education. In *36th International Workshop on Qualitative Reasoning*, 2023.

[19] B. Kuipers. Qualitative reasoning: modeling and simulation with incomplete knowledge. *Automatica*, 25(4):571–585, 1989.

[20] M. Law, A. Russo, and K. Broda. The ILASP system for learning answer set programs. www.ilasp.com, 2015.

[21] T. Miller. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, 36:e14, 2021.

[22] S. Muggleton. Inductive logic programming. *New Generation Computing*, 8:295–318, 1991.

[23] W. Pang and G. M. Coghill. Learning qualitative differential equation models: a survey of algorithms and applications. *The Knowledge Engineering Review*, 25(1):69–107, 2010.

[24] C. Preist, K. Eshghi, and B. Bertolino. Consistency-based and abductive diagnoses as generalised stable models. *Annals of Mathematics and Artificial Intelligence*, 11:51–74, 1994.

[25] R. Reiter. A theory of diagnosis from first principles. *Artificial intelligence*, 32(1):57–95, 1987.

[26] B. L. Richards, I. Kraan, and B. Kuipers. *Automatic abduction of qualitative models*. Citeseer, 1992.

[27] A. Sanfeliu and K.-S. Fu. A distance measure between attributed relational graphs for pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (3):353–362, 1983.

[28] A. C. Say and S. Kuru. Qualitative system identification: deriving structure from behavior. *Artificial Intelligence*, 83(1):75–141, 1996.

[29] C. Schut and B. Bredeweg. An overview of approaches to qualitative model construction. *The Knowledge Engineering Review*, 11(1):1–25, 1996.

[30] J. P. Smith III, A. A. DiSessa, and J. Roschelle. Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The journal of the learning sciences*, 3(2):115–163, 1994.

[31] B. C. Van Fraassen. *The scientific image*. Oxford University Press, 1980.