

Proceedings of the

QR 2024

37th International Workshop on Qualitative Reasoning

Co-located with the
European Conference on Artificial Intelligence (ECAI)
Santiago de Compostela – October 19th, 2024



Edited by

Diedrich Wolter¹ • Zoe Falomir²

¹ University of Lübeck, Germany

² University of Umeå, Sweden

Program Chairs

Zoe Falomir, Umeå University, Sweden

Diedrich Wolter, University of Lübeck, Germany

Program Committee

- Núria Agell, ESADE - Ramon Llull University, Spain
- Cecilio Angulo, Universitat Politècnica de Catalunya, Spain
- Ivan Bratko, University of Ljubljana, Slovenia
- Bert Bredeweg, University of Amsterdam, The Netherlands
- Stefano Borgo, ISTC CNR Istituto di Scienze e Tecnologie della Cognizione, Italy
- Maria Chang, IBM, USA
- Tony Cohn, University of Leeds, UK
- Vicent Costa, IIIA-CSIC, Spain
- Ken Forbus, Northwestern University, USA
- Joanna Golińska-Pilarek, University of Warsaw, Poland
- Florian Grigoleit, TU Munich, Germany
- Tomoya Horiguchi, Kobe University, Japan
- Liliana Ironi, IMATI - CNR, Italy
- Alexandra Kirsch, Tübingen, Germany
- Lleó Museros, University Jaume I, Spain
- Juan Carlos Nieves, Umeå University, Sweden
- Wei Pang, University of Aberdeen, Scotland, UK
- Kai-Florian Richter, Umeå University, Sweden
- Ismael Sanz, University Jaume I, Spain
- Quiang Shen, Aberystwyth University, Wales, UK
- Peter Struss, TU Munich, Germany
- Jacob Suchan, Constructor University Bremen, Germany
- Kazuko Takahashi, Kwansei Gakuin University, Japan
- Louise Travé-Massuyès, LAAS-CNRS, France
- Jure Žabkar, University of Ljubljana, Slovenia

Preface

Proceedings of the working papers accepted and presented at the 37th International Workshop on Qualitative Reasoning (QR) held on October 19th, 2024, co-located with ECAI in Santiago de Compostela, Spain. The workshop comprised discussions, presentations of technical papers, and two invited talks given by Nuriá Agell ("From Orders of Magnitude to Linguistic Nuance: Managerial Insights and Applications") and Vicent Costa ("Qualitative Modeling and Explainable Artificial Intelligence: Applications to Art Painting Style Categorization and Quality of Life for People with Disability").

The QR community is involved with the development and application of qualitative representations to understand the world from incomplete, imprecise, or uncertain data. Our qualitative models span natural systems (e.g., physics, biology, ecology, geology), social systems (e.g., economics, cultural decision-making), cognitive systems (e.g., conceptual learning, spatial reasoning, intelligent tutors, robotics), and more.

The QR community includes researchers in Artificial Intelligence, Engineering, Cognitive Science, Applied Mathematics, and Natural Sciences, commonly seeking to understand, develop, and exploit the ability to reason qualitatively. This broadly includes:

- Developing new formalisms and algorithms for QR.
- Building and evaluating predictive, prescriptive, diagnostic, or explanatory qualitative models in novel domains.
- Characterizing how humans learn and reason qualitatively about the (physical) world with incomplete knowledge.
- Developing novel, formal representations to describe central aspects of our world: time, space, change, uncertainty, causality, and continuity.

The International Workshop on Qualitative Reasoning provides a forum for researchers from multiple perspectives to share research progress toward these goals.

Topics of interest include:

- Qualitative modelling in physical, biological and social sciences, and in engineering.
- Representations and techniques for QR.

- Methods that integrate QR with other forms of knowledge representation, including quantitative methods, machine learning and other formalisms.
- Using QR for diagnosis, design, and monitoring of physical systems.
- Applications of QR, including education, science, and engineering.
- Cognitive models of QR, including the use of existing QR formalisms for cognitive modelling and results from other areas of cognitive science for QR.
- Using QR in understanding language, decision-making, sketches, images, and other kinds of signals and data sources.
- Formalization, axiomatization, and mathematical foundations of QR.

The accepted papers were reviewed by at least two members of the international program committee.

Acknowledgements

We acknowledge the support for organizing the Qualitative Reasoning workshop by the University of Lübeck, Germany and the University of Umeå, Sweden.

We thank the members of the international Program Committee for their valuable work during the reviewing process and we thank Nuriá Agell and Vicent Costa for their inspiring presentations. We also thank Bert Bredeweg for providing templates for the web site and these proceedings.

Contents

Qualitative Reference Model for Learning about Melatonin Regulation <i>Nihal Fawzi, Marco Kragten, and Bert Bredeweg</i>	6
Qualitative and quantitative modelling of dynamic systems: how do they relate? <i>Marco Kragten and Bert Bredeweg</i>	14
Reconstructing Qualitative Model Variations from Qualitative Descriptions for Conversational Explanation <i>Moritz Bayerkuhnlein and Diedrich Wolter</i>	23
Safe parking of a nonholonomic autonomous vehicle by qualitative reasoning <i>Domen Šoberl, Jan Lemeirec, Ruben Spolminkd, and Jure Žabka</i>	30
Using qualitative reasoning to compare media coverage of Israel Gaza war <i>Walaa Abuasaker, Núria Agell, Jennifer Nguyen, Nil Agell, Mónica Sánchez, and Francisco J. Ruiz</i>	36
Qualitative Modeling of Social Relationships <i>Ken Forbus</i>	40
Unveiling Ontological Commitment in Multi-Modal Foundation Models <i>Mert Kesera, Gesina Schwalbe, Niki Amini-Naienid, Matthias Rottmane, and Alois Knoll</i>	48
Preliminary Experiments of Qualitative Reasoning Model Construction Using Large Language Model <i>Shinpei Suzuki and Masaharu Yoshioka</i>	58
Using Qualitative Techniques with Kolmogorov-Arnold Networks for Explainable AI <i>Ismael Sanz, Lledó Museros, Vicente Casales-Garcia, and Luis Gonzalez-Abril</i>	64

Qualitative Reference Model for Learning about Melatonin Regulation

Nihal Fawzi¹, Marco Kragten¹ and Bert Bredeweg^{1,2}

¹ Faculty of Education, Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

² Informatics Institute, Faculty of Science, University of Amsterdam, Amsterdam, The Netherlands

{n.fawzi, m.kragten, b.bredeweg}@hva.nl

Abstract

Learning by creating qualitative representations is a valuable approach to learning. However, modelling is challenging for students, especially in secondary education. Support is needed to make this approach effective. To address this issue, we explore automated support provided to students while they create their qualitative representation. This support is generated from a reference model that functions as a norm. However, the construction of a reference model is still a challenge. In this paper, we present the reference model that we have created to support students in learning about the melatonin regulation in the context of the biological clock.

1 Introduction

Qualitative representations are used to aid students in learning about dynamic systems [1-5]. By creating a qualitative model, students actively develop their understanding of the subject matter as well as enhance their system thinking skills. However, qualitative representations are inherently complex and therefore difficult to construct [6,7]. Students require specific guidance that is detailed enough to facilitate progress, yet sufficiently restrained to leave ample room for discovery and learning [24]. To meet this need, we use reference models that function as a standard (as a norm) on the basis of which the necessary guidance is generated automatically [8,9].

We develop these reference models together with teachers and domain experts. Meanwhile, miscellaneous models have been created, particularly for usage in secondary education [10-14]. Yet, each new model remains a challenge, mainly for two reasons. Firstly, the source documentation is often incomplete, ambiguous and sometimes even contradictory which hampers extracting the relevant details and mapping them into the qualitative representation. Secondly, to act as a reference model for norm-based support (e.g., in secondary education) the model should adhere to specific requirements, including the following:

Graceful progression. The subject matter must be broken down into units, each representing a learnable yet adequately complex subsystem, while together these units are organized

into a logical sequence that incrementally encompasses the entire system.

Self-contained and manageable. Qualitative models can easily explode and generate large state-graphs, or conversely, not generate any states at all. To be suitable for learning, subsystems must generate simulations that provide correct solutions with for students manageable state-graph sizes.

Meaningful. The decomposition into units is not arbitrary, on the contrary, each subsystem should by itself address at least one, possibly a few, important, meaningful, and valuable features of the subject matter.

Intriguing and curiosity driven. Surprises may help stimulate students' curiosity and their drive to wanting to address the next challenge [23]. Simulation results can be used for this. Hence, the goal is to orchestrate modelling steps such that when simulated they regularly produce intriguing results, which then become the challenge to be addressed in the next modelling step.

In this paper, we present the reference model that we developed for aiding students in learning about melatonin regulation, as well as the decomposition of this model into a sequence of learnable units. Melatonin is a hormone that is part of the mechanism that regulates the 24-hour rhythm of the biological clock. Understanding the biological clock and how it impacts life is in principle universally relevant yet typically not deeply embedded in formal education (at least not in the Netherlands). As such, the biological clock presents an interesting and relevant case.

The content of this paper is as follows. Section 2 summarizes the subject matter. Section 3 describes the qualitative representation software we use. Section 4 presents the reference model, with subsections for each mechanism from the full system. Section 5 and 6 conclude the paper.

2 Biological clock

The biological clock, also known as the circadian clock, is a cycle that takes place roughly within 24 hours. It is an autonomous series of responses in biological species that synchronizes with the day-night cycle. Before humans invented candles and the use of electricity for lamps, organisms relied on natural light only, resulting in the internal clock being in sync with this cycle. In modern times, however, the biological clock is disrupted by the 24-hour society in which people live. Research shows this disruption

has a major impact on human, animal and plant behaviour, as well as on whole ecosystems [15,16].

For the work presented here we focus on a particular aspect of the human circadian clock, namely the hormone melatonin and how its concentration changes during 24 hours.

The main driver of the biological clock is a group of nerve cells called the Suprachiasmatic Nucleus (SCN). The SCN inhibits the inhibiting work of the Paraventricular Nucleus

(PVN). This results in more Aralkylamine N-acetyltransferase (AANAT), because less AANAT is being degraded. Next, because AANAT drives the conversion of serotonin into melatonin, the latter now increases. This process has a cyclic nature, because the initial cause, the SCN, has a cyclic behaviour. Fig. 1 illustrates the mechanism in more detail.

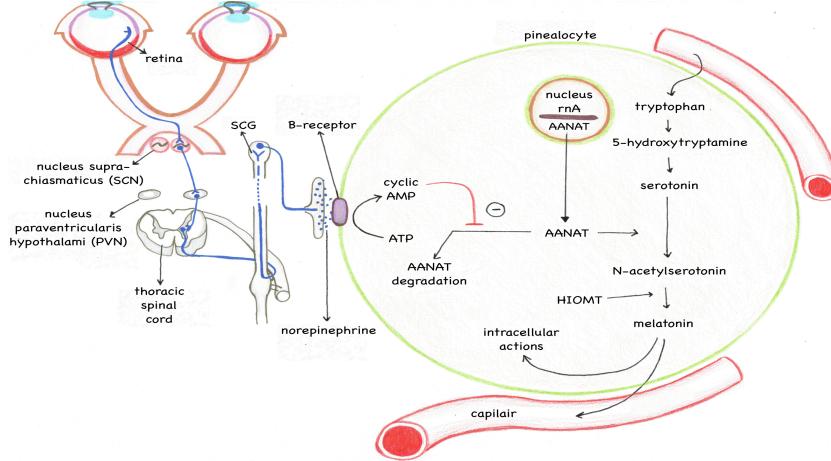


Fig. 1. Artist impression of the biological clock mechanism according to [17] (but see also [18]).

3 Qualitative Reasoning with DynaLearn

DynaLearn (<https://www.dynalearn.nl>) is an interactive tool that allows learners to create and simulate qualitative representations. It provides a web-based graphical user interface to Garp3 [19], facilitating online usage of the latter. The following ingredients are available via this interface to create representations. *Entities* can be used for representing physical objects and/or abstract concepts that make up the system. *Configurations* can be used for representing structural relationships between entities. *Quantities* can be used for representing changeable and measurable features of entities. Quantities have *Direction of change* (∂) (decreasing, steady, and increasing) and a *Quantity space* (a set of alternating point and interval values that the quantity can take on). *Causal dependencies* can be used for representing directed relationships between quantities. *Correspondences* can be used for representing co-occurring values and co-occurring directions of change. *In/equalities* can be used for representing order information among values and among directions of change. Finally, there is the option to represent *conditional statements*: IF A THEN B , where A and B can refer to the ingredients mentioned above.

When simulating, *Initial values* are defined for quantities, typically (but not exclusively) at the start of *Causal paths* (sequences of causal dependencies). This can be a direction of change, an initial value or an *Exogenous* behaviour. Additionally, in/equalities can be specified.

The simulation produces a *State-graph*, which consist of one or more *States* (unique qualitative behaviour of the system) and possibly *Transitions* (continuous passage) between pairs of states. The changes of system behaviour

throughout the state-graph can be inspected using the *Value-history* and the *Inequality-history*.

Introducing advanced tooling in education requires a stepwise approach regarding complexity. To accommodate this, DynaLearn can be used at different levels of complexity [20]. The ideas presented in this paper are situated at level 4, which encompasses a large set of available ingredients. Importantly, this level includes the causal dependencies influence (I^+/I^-) and proportionality (P^+/P^-) [21]. Learners can thus focus on the distinction between processes (I) (initial causes) and the propagation (P) of these through the system. Positive and negative feedback loops are also available and in/equality ($< \leq = \geq >$) can be used to represent the relative impact of competing processes.

4 The Reference Model

The final version of the reference model, as we developed it, is shown in Fig. 13. We first developed the complete model, based on the required learning goals, and after that decomposed this model into units for learning.

The complete model can roughly be divided into four subsystems: the pulse generation (left), the AANAT regulation (middle), the production of serotonin (right-top), and the production of melatonin (right-bottom). Table 2 shows the simulation results focussing on the current value and direction of change for each quantity in each state. For instance, *AANAT Degradation* reaches its highest point and becomes momentarily steady in state 7 ($\langle M, 0 \rangle$) and then starts decreasing in state 8 ($\langle M, - \rangle$). Table 3 shows the inequality information. For instance, *AANAT Degradation* and *Production* are equal in state 4 ($=$), while *Degradation* has become higher in state 5 ($>$). Notice that the behaviour of

the system is cyclic. The state transitions follow the path as shown by the state-graph in Fig. 13.

4.1 Production and degradation of AANAT

It may seem logical to start the learning with the initial change at the start of the causal path. However, starting with the production and degradation of the *AANAT* is preferred. The main reason being that this combination of processes is the richest place in the whole model, with many opportunities for introducing key notions of systems thinking combined with domain knowledge. This is achieved without the added complexity of an oscillating impulse (see Fig. 13), which could lead to many states that are not yet useful for learning about this part of the mechanism.

The instruction for the lesson is given to the students via a workbook (on paper). The first assignments in the workbook thus focusses on modelling the production of *AANAT* (Fig. 2). This entity must be given two quantities: *Amount* and *Production*. The latter has a positive influence on the former ($I+$), while the production itself remains steady due to the exogenous influence. An influence requires a quantity space, here $\{0, +, \text{Max}\}$, because we need to know if the causing ‘value’ is positive or negative. In Fig. 2, this value is 0 and not causing any effect. Hence, when simulating, *Amount* remains steady. From this point, the workbook moves to the details in Fig 3, in which the value of the influencing quantity (*Production*) is set to $+$. Now the influence does cause an effect and hence *Amount* is increasing.

At this point in the lesson, the student is required to vary model details, run simulations, and answer questions about the results. The workbook instructions guide these steps.

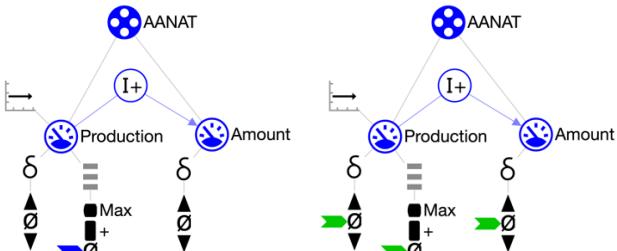


Fig. 2. AANAT production. Left side shows the model with production initially being set to 0 and steady (due to the exogenous influence \sqsubseteq). Right side shows the simulation results. Because the process is inactive, nothing changes.

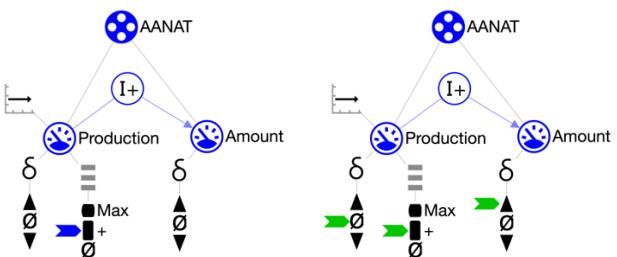


Fig. 3. AANAT production. Left side shows the model with production initially being set to $+$ and steady (due to the exogenous influence \sqsubseteq). Right side shows the simulation results. Because the process is active, the amount of AANAT increases, while the process itself remains steady.

After production is sufficiently addressed, the next step is to add degradation as a competing process. Fig. 4 shows the result. *Degradation* has a positive current value (+) and a negative influence on *Amount* ($I-$). However, only specifying this information is insufficient, resulting in an ambiguous simulation with miscellaneous solutions. For instance, quantity spaces are (by definition) independent sets of ordered values, with only 0 as a universal. Hence, in Fig 4 the values *Max* for *Production* (P_{Max}) and for *Degradation* (D_{Max}) are unrelated, and all options are in principle valid (thus: $P_{\text{max}} < D_{\text{Max}}$, $P_{\text{max}} = D_{\text{Max}}$, $P_{\text{max}} > D_{\text{Max}}$) unless more information is specified. A similar situation holds for the balance between the *Production* and *Degradation* processes, all options are possible (thus: $P < D$, $P = D$, $P > D$).

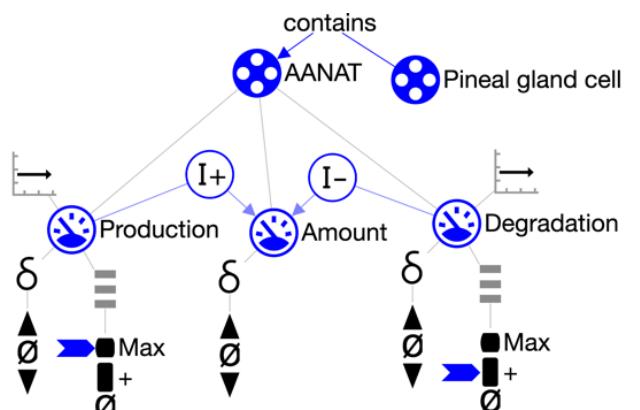


Fig. 4. AANAT production and degradation.

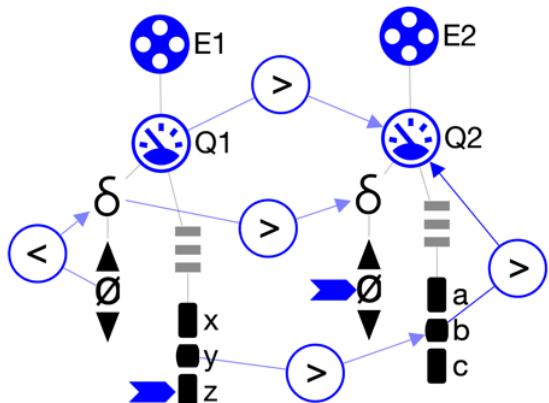


Fig. 5. Inequality information in a qualitative model.

Although ambiguity in qualitative models is typically considered to be a burden, here it provides an opportunity to intrigue students and stimulate them to further refine their understanding of the system. Let’s consider an example. Fig. 5 illustrates the kinds of inequality information and their role for describing unique characteristics of systems. Let’s assume that quantities Q_1 and Q_2 refer to the mutual temperatures (T) of the entities E_1 (T_{E1}) and E_2 (T_{E2}), respectively. The details in Fig 5 can then be read as follows:

- Current value of T_{E1} is z ($Q_1 = z$).
- T_{E1} is increasing ($0 < \partial Q_1$).

- Current value of T_{E2} is below b ($b > Q_2$).
- T_{E2} is steady ($\partial Q_2 = 0$).
- Current value of T_{E1} is greater than the current value of T_{E2} ($Q_1 > Q_2$).
- T_{E1} is increasing faster than T_{E2} ($\partial Q_1 > \partial Q_2$).
- If we assume that y and b are boiling points of Q_1 and Q_2 , respectively, then the boiling point of Q_1 is higher than the boiling point of Q_2 ($y > b$).

When building a model, students must think explicitly about such system details and figure out the appropriate facts.

With the above-mentioned options in mind, the workbook continues by prompting students to think about additional (relevant) details regarding the two processes influencing *AANAT*. Fig 6 shows a particular situation in this context. In comparison to Fig. 4, it is now also known that the highest possible level of *Production* equals that of the highest possible level of *Degradation* ($P_{\max} = D_{\max}$). It is also known that currently *Production* is higher than *Degradation* ($P > D$). Simulating this model delivers a state-graph with four consecutive states: $1 \rightarrow 2 \rightarrow 3 \rightarrow 4$. Table 1 summarises the results. It shows that there is a steady *Production* in all states. *Degradation* on the other hand is increasing. Initially it is smaller than *Degradation* (S_1), than it becomes equal (S_2), and finally it outperforms *Production* (S_3 and S_4). Due to this, *Amount* initially increases (S_1), becomes steady (S_2), and then decreases (S_3 and S_4).

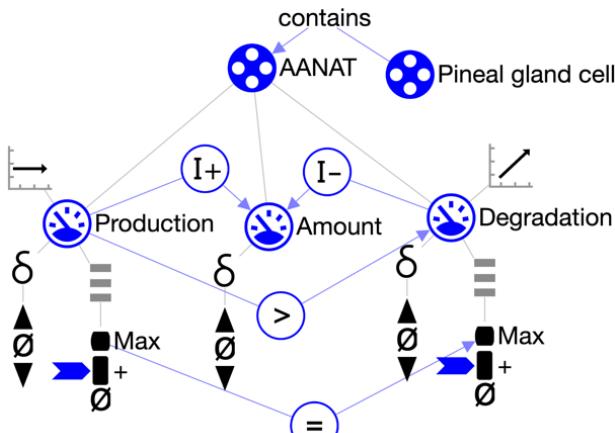


Fig. 6. Additional information regarding the two processes influencing AANAT.

Table 1. Simulation results for the model shown in Fig. 6. S refers to State, M to Max, P to Production, A to Amount, D to Degradation, and u refers to unspecified value.

	S_1	S_2	S_3	S_4
P	$<+, 0>$	$<+, 0>$	$<+, 0>$	$<+, 0>$
A	$<u, +>$	$<u, 0>$	$<u, ->$	$<u, ->$
D	$<+, +>$	$<+, +>$	$<+, +>$	$<M, 0>$
(P ? D)	$P > D$	$P = D$	$P < D$	$P < D$

To conclude this part of the lesson, the workbook asks students to draw line-graphs (on paper) of how the quantities change over time. Fig. 7 shows a graph they must complete.

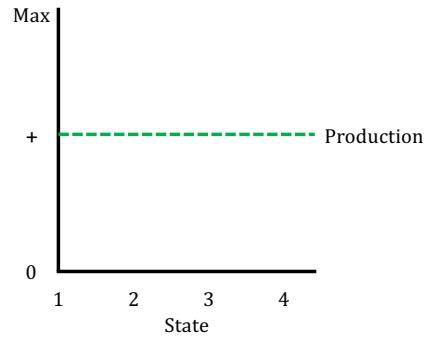


Fig. 7. Student assignment (on paper): Draw the line-graph of how AANAT changes according to the simulation results.

4.2 Regulation of AANAT degradation

The next logical step in the model is to focus on the mechanism that controls the *AANAT* degradation process. Why? Part of the reason is that correct behaviour of *AANAT* is a prerequisite before being able to discuss the other effects that follow and are subsequently controlled by *AANAT*.

The workbook introduces the topic with the visual and the textual explanation shown in Fig. 8. Notice that the *SCN* influences the *Degradation* process via a double negation including the *PVN*. Hence, the *Degradation* process follows the *SCN* rhythm. Students find a double negation in a causal chain sometimes difficult.

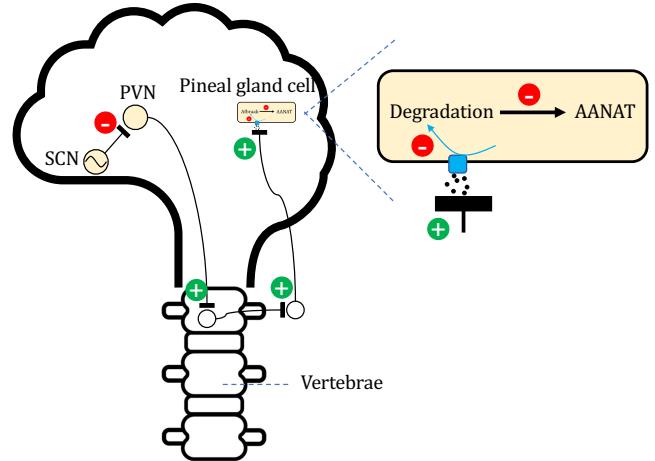


Fig. 8. Workbook information source (drawing & text). Text: ‘The SCN (suprachiasmatic nucleus) is an area of the brain that contains clock genes that determine the 24-hour rhythm of many processes in the body by sending out impulses. The SCN has an inhibitory effect on the PVN (paraventricular nucleus). Through several nerve cells, the PVN has an inhibitory effect on the breakdown of AANAT.’

Fig. 9 shows the qualitative representation. The *SCN* impulses have a negative proportional influence on the *PVN* impulses, which in turn has a negative proportional influence on the *Degradation* process. By choice, some intermediate causal dependencies are not included in the representation. The *SCN* quantity *Impulse* is given an exogenous starting behaviour (type: sinusoidal) [22]. This implements the

sinusoidal behaviour of the internal clock. To ensure that *Degradation* fully follows the *SCN*, a quantity space correspondence (*C*) between the two quantity spaces is needed. Note that a quantity space for the *SCN* quantity *Impulse* is strictly speaking not needed. However, adding it makes the sinusoidal behaviour more visible as during the sequence the quantity now changes values.

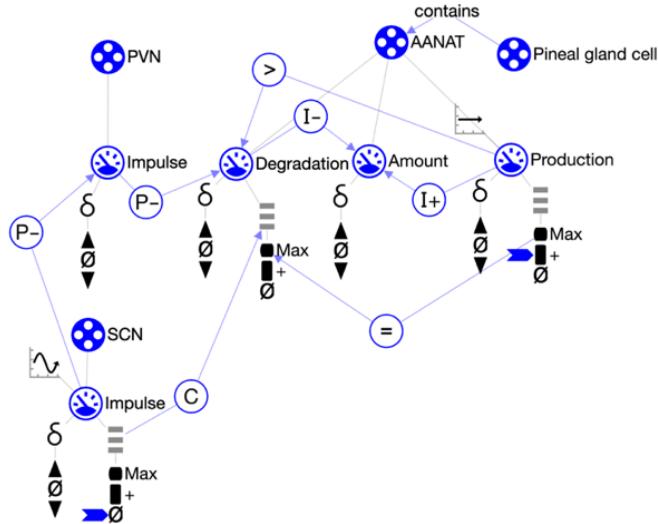


Fig. 9. AANAT production as shown in Fig 6, augmented with the SCN (and PVN) which controls the degradation process.

Fig. 10 and 11 show the simulation results. The state-graph has 10 consecutive states (Fig. 10). The *SCN Impulse* has a cycle behaviour, due to the exogenous influence. It starts at value 0 in state 1 <0, 0> (Fig. 11), increases to value Max in state 6 <Max, 0>, starts decreasing again in state 7 <Max, ->, and via state 10 <+, -> goes back to 0 and steady in state 1. The *PVN Impulse* changes opposite from this. It is also momentary steady in state 1, but then it decreases in states 2 to 5, becomes momentary steady in state 6 and increases in states 7 to 10. The *AANAT Degradation* behaves opposite from the *PVN* and hence follows the original *SCN* behaviour. *AANAT Production* is not shown in Fig. 11, but from Fig. 9 we can see that it has value + and remains steady due to an exogenous influence, hence <+, 0> in all states. Because *AANAT Degradation* changes, the balance between *AANAT Production* and *Degradation* varies over the consecutive states. This is shown in the inequality history in Fig. 11 (bottom). In state 1 to 3 *Production* dominates and *AANAT Amount* increases. In state 4 the two processes reach a balance and *AANAT Amount* stops increasing. In state 5 to 8 *Degradation* dominates and *AANAT Amount* decreases. In state 9 the two processes balance again and *AANAT Amount* stops decreasing. The *Amount* increases again in state 10.

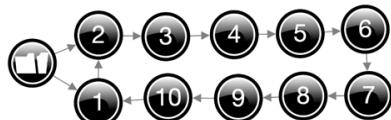


Fig. 10. State-graph when simulating the representation in Fig. 9.

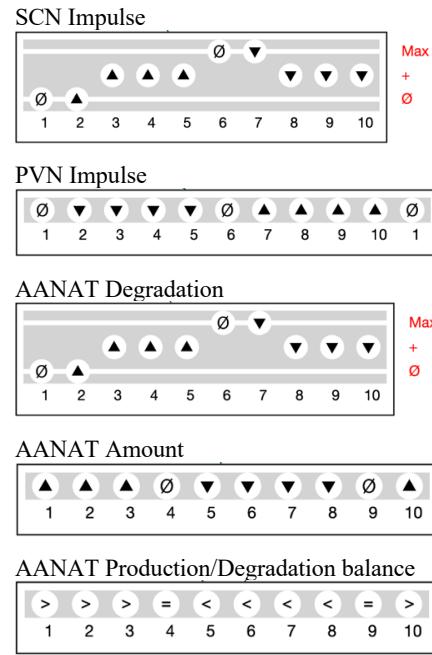


Fig. 11. Simulation results of the SCN controlling the AANAT as shown in the representation in Fig. 9.

4.3 Conversion and degradation of melatonin

Melatonin is produced in the pineal gland cells and then goes to the blood. The liver breaks it down again. The representation details are shown in Fig 12. It continues with adding the entities *Melatonin* and *Liver* (blood is not modelled). *Conversion* to the *Pineal gland cells*, *Degradation* to *Liver*, and *Amount* to *Melatonin*. The *Conversion* process is proportional to *AANAT Amount*, while *Conversion* and *Degradation* each influence *Melatonin Amount*. By placing an inequality, we can track the balance between them. Finally, there is negative feedback from *Melatonin Amount* on *Degradation* (*P+*).

The simulation now produces 12 states, similar to (in fact a subset of) the details shown in Fig. 13 and Table 2 and 3.

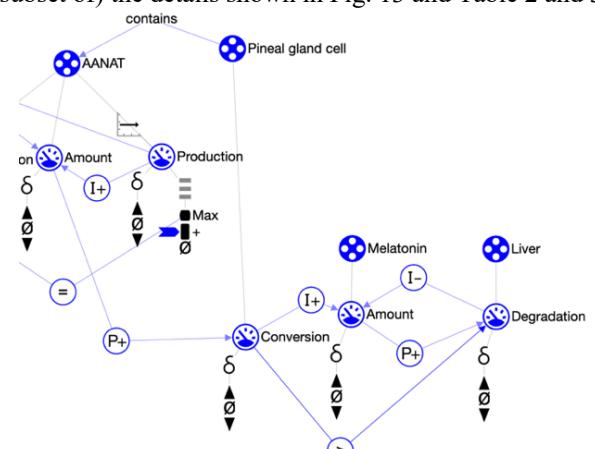


Fig. 12. Conversion and degradation of melatonin added to the representation shown in Fig 9. Note, to maintain readability we cropped the figure. See Fig. 9 for the remaining context.

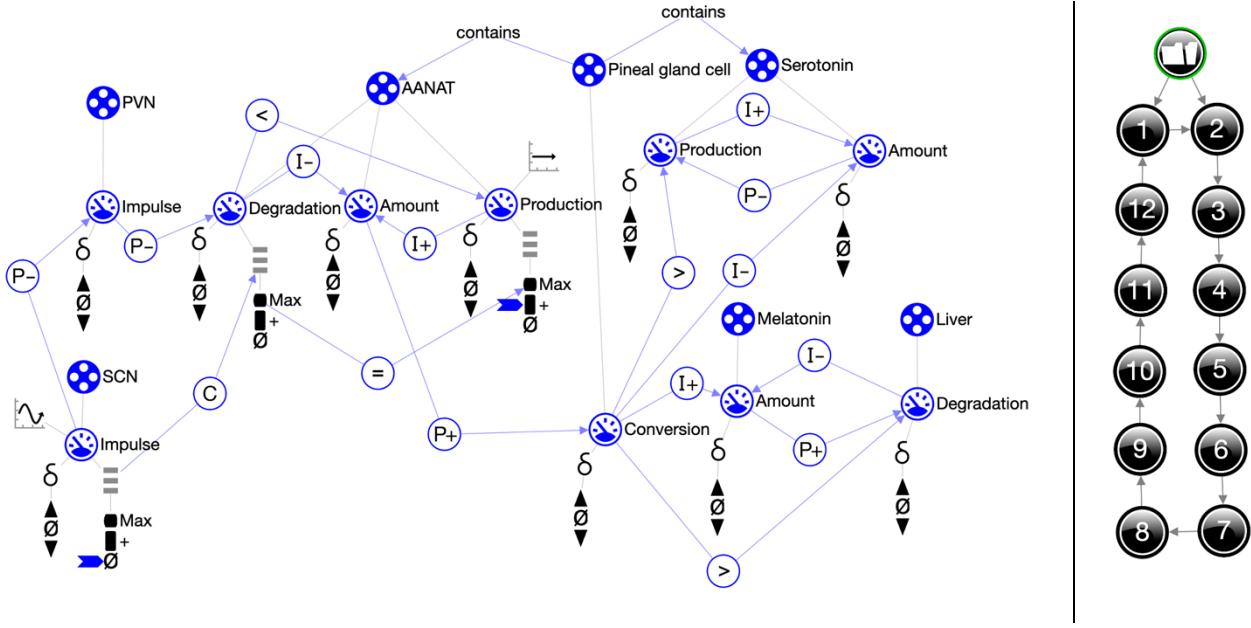


Figure 13. Qualitative reference model of melatonin regulation. Left side shows the representation. Right side shows the simulation results as a state-graph which consists of a loop of 12 consecutive states.

Table 2. Simulation results for the melatonin regulation reference model shown in Figure 13. PGC refers to Pineal gland cell, S refers to State, $\langle v, \partial v \rangle$ refers to value and derivative (change), respectively, M refers to Max, and u refers to unspecified value.

Entity	Quantity	S ₁	S ₂	S ₃	S ₄	S ₅	S ₆	S ₇	S ₈	S ₉	S ₁₀	S ₁₁	S ₁₂
AANAT	Amount	$\langle u, + \rangle$	$\langle u, + \rangle$	$\langle u, + \rangle$	$\langle u, 0 \rangle$	$\langle u, - \rangle$	$\langle u, 0 \rangle$	$\langle u, + \rangle$	$\langle u, + \rangle$				
	Production	$\langle +, 0 \rangle$											
	Degradation	$\langle 0, 0 \rangle$	$\langle 0, + \rangle$	$\langle +, + \rangle$	$\langle M, 0 \rangle$	$\langle M, - \rangle$	$\langle +, - \rangle$						
Liver	Degradation	$\langle u, + \rangle$	$\langle u, + \rangle$	$\langle u, + \rangle$	$\langle u, 0 \rangle$	$\langle u, 0 \rangle$	$\langle u, - \rangle$	$\langle u, 0 \rangle$	$\langle u, 0 \rangle$	$\langle u, + \rangle$			
Melatonin	Amount	$\langle u, + \rangle$	$\langle u, + \rangle$	$\langle u, + \rangle$	$\langle u, 0 \rangle$	$\langle u, 0 \rangle$	$\langle u, - \rangle$	$\langle u, 0 \rangle$	$\langle u, 0 \rangle$	$\langle u, + \rangle$			
PGC	Conversion	$\langle u, + \rangle$	$\langle u, + \rangle$	$\langle u, + \rangle$	$\langle u, 0 \rangle$	$\langle u, - \rangle$	$\langle u, 0 \rangle$	$\langle u, + \rangle$	$\langle u, + \rangle$				
PVN	Impulse	$\langle u, 0 \rangle$	$\langle u, - \rangle$	$\langle u, 0 \rangle$	$\langle u, + \rangle$								
SCN	Impulse	$\langle 0, 0 \rangle$	$\langle 0, + \rangle$	$\langle +, + \rangle$	$\langle +, + \rangle$	$\langle +, + \rangle$	$\langle M, 0 \rangle$	$\langle M, - \rangle$	$\langle +, - \rangle$				
Serotonin	Amount	$\langle u, - \rangle$	$\langle u, - \rangle$	$\langle u, - \rangle$	$\langle u, 0 \rangle$	$\langle u, 0 \rangle$	$\langle u, + \rangle$	$\langle u, + \rangle$	$\langle u, + \rangle$	$\langle u, 0 \rangle$	$\langle u, 0 \rangle$	$\langle u, - \rangle$	$\langle u, - \rangle$
	Production	$\langle u, + \rangle$	$\langle u, + \rangle$	$\langle u, + \rangle$	$\langle u, 0 \rangle$	$\langle u, 0 \rangle$	$\langle u, - \rangle$	$\langle u, - \rangle$	$\langle u, - \rangle$	$\langle u, 0 \rangle$	$\langle u, 0 \rangle$	$\langle u, 0 \rangle$	$\langle u, + \rangle$

Table 3. Simulation results cont. showing the inequality information for three quantity pairs in each of the states.

Compared quantities	S ₁	S ₂	S ₃	S ₄	S ₅	S ₆	S ₇	S ₈	S ₉	S ₁₀	S ₁₁	S ₁₂
AANAT: Degradation versus AANAT: Production	<	<	<	=	>	>	>	>	=	<	<	<
Pineal gland cell: Conversion versus Serotonin: Production	>	>	>	=	=	<	<	<	<	=	=	>
Pineal gland cell: Conversion versus Liver: Degradation	>	>	>	=	=	<	<	<	<	=	=	>

4.4 Conversion and production of serotonin

The final part of the model concerns the production and conversion of serotonin. The details are show in Fig 13 (right hand top). Serotonin is produced (from tryptophan) and then converted to melatonin using AANAT. In the representation this is slightly simplified making details kind of analogous to the mechanism for melatonin.

The *Conversion* by the *Pineal gland cells*, negatively influences the *Serotonin Amount*, because it is used to create the melatonin. The *Serotonin Production* is negative

proportional to the *Serotonin Amount* (together implementing a negative feedback loop). The inequality between the *Pineal gland cells Conversion* and the *Serotonin Production* is not needed for arriving at the correct simulation results, but it helps to make the balance between these two processes visible, and thereby the mechanism potentially more insightful for students.

Note that there is no feedback from the amount of serotonin and the amount of melatonin on the conversion process (*Pineal gland cell Conversion*). This feedback was

not included for two reasons. First, the main driver for the conversion is *AANAT Amount*. Second, such feedback loops result in extra states making the simulation harder to interpret. Adding additional information to circumvent those extra behaviours requires adding more ingredients in the representation, which would also make the lesson more complex. Together that lead to the decision to not include this feedback.

The representation is now complete. The simulation produces the results as shown in Fig. 13 and Table 2 and 3. To conclude the lesson, the workbook asks students to draw a line-graph (on paper) showing the changes of Serotonin during a full 24-hour cycle. Fig. 14 shows the graph they must complete.

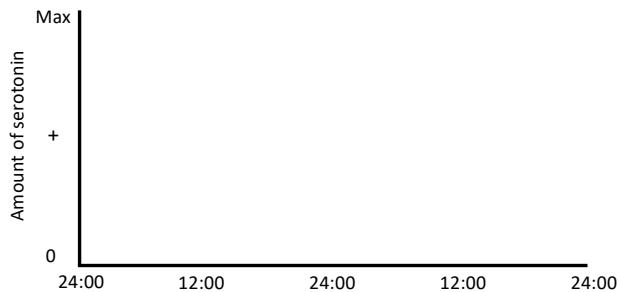


Figure 14. Student assignment (on paper): Draw a line-graph showing how the amount of serotonin changes according to the simulation results.

5 Towards evaluation

The reference model presented in this paper is part of our research effort to make ‘learning by building qualitative models’ a valuable approach. For this purpose, 8 students (upper secondary education), 5 teachers in training (higher education) and 4 teacher educators (higher education) have taken the lesson based on the model presented in this paper.

These users were all novices in the sense that they had no previous experience with qualitative modelling. On average they took 110 minutes to complete the lesson. Except for one subject, they all completed the lesson in the allotted time. Pre- and post-tests suggest learning effects for Systems thinking (from 6.4 to 12.4 out of 16 points) and for Melatonin regulation (from 4.8 to 8.8 out of 16 points), but these results may also indicate that the learning experience can be improved (although there is a limit to what can be learned in two hours). The users seem to have enjoyed the lesson, as they graded their experience with an 8 (on a scale of 10). However, these results are all preliminary. These lessons have been recorded and are currently being analysed to investigate the (*i*) support *use* and (*ii*) support *need* that these users have. The results will be input to further advance the learning by modelling approach.

The development of the reference model described in this paper underwent several improvements before reaching its final status. Critical expert reviews were provided by researchers who have published scientific justifications of the mechanisms (cf. [17,18]) to ensure that the model reflects the latest scientific insights on the topic.

6 Conclusion and Discussion

Reference models are an important asset in our approach to support students in learning from creating qualitative models. In this paper, we present a reference model for learning about melatonin regulation. Melatonin regulation is a particular aspect of the human circadian clock (also known as the biological clock). The model was developed in close collaboration with subject matter experts to ensure validity.

The model has four interacting processes, which together cause a serious amount of ambiguity upon simulating, easily resulting in complex state-graph consisting of 90 qualitative states with many alternative paths between those states. The shortest path algorithm hides alternative paths between two states leaving the shortest path (while ensuring certain constraints to maintain correct results [22]). By deploying the fastest path heuristic, the simulation results are simplified leaving only the behaviour relevant to explaining the quintessence of the regulation mechanism.

After development, the model was disentangled into four units that together form the system: (*i*) pulse generation, (*ii*) AANAT regulation, (*iii*) production of serotonin, and (*iv*) production of melatonin. The decomposition followed guidelines to guarantee learnability, specifying that units should facilitate (*i*) graceful progression, as well as being (*ii*) self-contained and manageable, (*iii*) meaningful, and (*iv*) intriguing and curiosity driven. Next, the order of the units in the overall assignment was arranged such that learning experience per unit was maximised as much as possible.

Future research focusses on advancing our automated support to aid students in learning from building qualitative models. For that purpose, seventeen users have taken the lesson build on the reference model presented in this paper. These data are currently being analysed.

Creating valuable qualitative models is still cumbersome. It requires a significant amount of craftsmanship based on experience. Future research could focus on automating this knowledge engineering endeavour and create tools that make building such models easier.

Acknowledgments

The research presented here is part of the BioClock Consortium which is funded by the NWA-ORC programme of the Dutch Research Council (NWO), project number 1292.19.077, <https://bioclockconsortium.org/>. We would like to thank Marijke Gordijn and Laura Kervezee for their valuable feedback on the model, as well as the students, teachers in training, and teacher educators for helping us evaluating the lesson.

7 References

- [1] Bredeweg, B. & Forbus. K.D. (2003). Qualitative Modeling in Education. *AI Magazine*, 24(4), 35-46.
- [2] Bredeweg, B. & Forbus. K.D (2016). Qualitative Representations for Education. In: R. A. Sottilare, A. C. Graesser, X. Hu, A. M. Olney, B. D. Nye, & A. M.

- Sinatra (Eds.), *Design Recommendations for Intelligent Tutoring Systems: Domain Modeling*, 4, 55-68.
- [3] Bredeweg, B., Liem, J., Beek, W., Linnebank, F., Gracia, J., Lozano, E., Wißner, M., Bühlung, R., Salles, P., Noble, R., Zitek, A., Borisova, P. & Mioduser, D. (2013). DynaLearn - An intelligent learning environment for learning conceptual knowledge. *AI Magazine*, 34(4), 46-65.
 - [3] Gautam Biswas, G., Segedy J.R. & Bunchongchit, K. (2016). From Design to Implementation to Practice a Learning by Teaching System: Betty's Brain. *International Journal of Artificial Intelligence in Education*, 26, 350-364.
 - [4] Forbus, K.D., Whalley, P.B., Everett, J.O., Ureel, L., Brokowski, M., Baher, J. & Kuehne, S.E. (1999). CyclePad: An articulate virtual laboratory for engineering thermodynamics. *Artificial Intelligence*, 114, 297-347.
 - [5] Forbus, K.D., Carney, K.E., Sherin, B. & Ureel, L. (2004). VModel: A visual qualitative modeling environment for middle-school students. Proceedings of the 16th Innovative Applications of Artificial Intelligence Conference, San Jose, USA.
 - [6] Jochem Liem, J. (2013). Supporting conceptual modelling of dynamic systems: A knowledge engineering perspective on qualitative reasoning. PhD thesis, University of Amsterdam, The Netherlands.
 - [7] Lozano, E., Gracia, J., Corcho, O., Noble, R.A. & Gómez-Pérez. (2015). A Problem-based learning supported by semantic techniques, *Interactive Learning Environments*, 23(1), 37-54, 2015.
 - [8] Bredeweg, B., Kragten, M., Holt, J., Kruit, P., van Eijck, T., Pijs, M., Bouwer, A., Sprinkhuizen, M., Jaspar, E., & de Boer, M. (2023). Learning with Interactive Knowledge Representations. *Applied Sciences*, 13(9), Article 5256.
 - [9] Bredeweg, B., Kragten, M., & Spitz, L. (2021). Qualitative Representations for Systems Thinking in Secondary Education. Paper presented at 34th International Workshop on Qualitative Reasoning, Montreal, Canada.
 - [10] Spitz, L., Kragten, M., & Bredeweg, B. (2021). Learning Domain Knowledge and Systems Thinking using Qualitative Representations in Secondary Education (grade 8-9). 34th International Workshop on Qualitative Reasoning, Montreal, Canada.
 - [11] Kragten, M., Spitz, L., & Bredeweg, B. (2021). Learning Domain Knowledge and Systems Thinking using Qualitative Representations in Secondary Education (grade 9-10). 34th International Workshop on Qualitative Reasoning, Montreal, Canada.
 - [12] Kragten, M., Jaspar, E. J. O. A., & Bredeweg, B. (2022). Learning Domain Knowledge and Systems Thinking using Qualitative Representations in Secondary Education (grade 10-12). 35th International Workshop on Qualitative Reasoning, Vienna, Austria.
 - [13] Kragten, M., Hoogma, T., & Bredeweg, B. (2023). Learning domain knowledge and systems thinking using qualitative representations in upper secondary and higher education. 36th International Workshop on Qualitative Reasoning, Krakow, Poland.
 - [14] Kragten, M., & Bredeweg, B. (2023). Describing the characteristics of circular and elliptical motion using qualitative representations. 36th International Workshop on Qualitative Reasoning, Krakow, Poland.
 - [15] Kumar, V. (ed.). (2017). Biological Timekeeping: Clocks, Rhythms and Behaviour. Springer, Delhi, India.
 - [16] Philips, A. et al., (2019). High sensitivity and interindividual variability in the response of the human circadian system to evening light. *PNAS*. 116(24), 12019-12024.
 - [17] Gordijn, M.C.M., (2018). Melatoninebehandeling voor slaap-waakstoornissen. *Psyfar*, 1, 9-17.
 - [18] Benarroch, E., (2008). Suprachiasmatic nucleus and melatonin. *Neurology*; 71, 594-598.
 - [19] Bredeweg, B., Linnebank, F., Bouwer, A. & Liem, J. (2009). Garp3 - Workbench for qualitative modelling and simulation. *Ecological Informatics*, 4(5-6), 263-281.
 - [20] Bredeweg, B., Liem, L., Beek, W., Salles, P. & Linnebank, F. (2010). Learning spaces as representational scaffolds for learning conceptual knowledge of system behaviour. In: M. Wolpers, P. A. Kirschner, M. Scheffel, S. Lindstaedt, & V. Dimitrova (Eds.), *Technology Enhanced Learning*, LNCS 6383, 46-61.
 - [21] Forbus, K.D. (2018). Qualitative representations. How people reason and learn about the continuous world. Cambridge, Massachusetts: The MIT Press.
 - [22] Bredeweg, B. & Linnebank, F.E. (2012). Simulation preferences - Means towards usable QR engines. 26th International Workshop on Qualitative Reasoning, Playa del Rey, CA, USA.
 - [23] Renninger, K.A. & Hidi, S.E. (2019). The Cambridge Handbook of Motivation and Learning. Cambridge University Press.
 - [24] Marco Kragten, M. & Bredeweg, B. (2024). Calcium Regulation Assignment: Alternative Styles in Successfully Learning about Biological Mechanisms. In: Artificial Intelligence in Education, LNAI 14829, 220-234. Springer, Cham.

Qualitative and quantitative modelling of dynamic systems: how do they relate?

Marco Kragten¹ and Bert Bredeweg^{1,2}

¹Faculty of Education, Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

²Informatics Institute, Faculty of Science, University of Amsterdam, Amsterdam, The Netherlands

{m.kragten, b.bredeweg}@hva.nl

Abstract

In this paper, we focus on how the qualitative vocabulary of Dynalearn, which is used for describing dynamic systems, corresponds to the mathematical equations used in quantitative modeling. Then, we demonstrate the translation of a qualitative model into a quantitative model, using the example of an object falling with air resistance.

1 Introduction

Understanding the behaviour of dynamic systems (e.g., climate change, economic growth and recession, population dynamics) is an important goal in secondary education. Educational developments that strive for future-oriented curricula emphasize this and consider practices such as causal reasoning and modelling as important skills.

Modelling is widely recommended as a way to provide learners with a deeper understanding of dynamic systems [1]. Modelling of a dynamic system on the computer can be done both qualitatively and quantitatively. Both forms can and are used in education [2, 3], but largely independent of each other. Both forms of modelling have their unique ways of representing and reasoning about system behaviour. As learning tools, each has its own pedagogical approach and offers distinct advantages and downfalls for understanding systems [4, 5, 6, 7]. Quantitative modelling allows for precise predictions and is closely aligned with the content of various school subjects such as gravitational acceleration (physics), predator-prey relationships (biology), the pig cycle (economics), and global warming (geography). Qualitative modelling, on the other hand, aligns more closely with the human reasoning about systems and emphasizes causality and the potential states of a system [8]. It also allows for automated support [2].

Education will benefit from a software solution and corresponding pedagogical approach that supports the strengths of both modelling forms. The software should integrate qualitative and quantitative representations of a system. If learners construct a qualitative model, the software can assist in translating it into a quantitative model, which learners often find challenging. Conversely, moving from

quantitative to qualitative helps to verify whether the constructed quantitative model assumes plausible causal relationships. This approach also aligns with recommendations from the scientific community [9]. It is important to note that such software does not yet exist, and that the potential impact of this innovation could extend to many other sectors in society.

In this paper, we focus on how qualitative representations of dynamic systems in Dynalearn [10] relate to mathematical equations. Chapter 2 begins by outlining the qualitative vocabulary of Dynalearn. We then discuss in Chapter 3 how dynamic systems can be quantitatively described using mathematical equations. A considerable portion of this paper, Chapter 4, is dedicated to examining the relationships between the qualitative vocabulary of Dynalearn and the corresponding mathematical equations. Following this, we use the dynamics of an object falling with air resistance as a case study in Chapter 5 to demonstrate the translation of a qualitative model into a quantitative model. The paper finalizes with a conclusion and discussion in Chapter 6.

2 Qualitative modelling

Qualitative representations provide a framework for modelling dynamic systems without relying on numerical data. The Dynalearn software facilitates the construction of these models at five distinct levels of complexity, each introducing new ingredients to accommodate a more nuanced description of system dynamics. In this paper we focus on level 4. Hence, this section discusses the ingredients of the Dynalearn software at that level.

Entities are either physical objects or abstract concepts, characterized by one or more *quantities*—changeable features of entities, such as temperature or speed. Each quantity has a *derivative*, denoted as δ , indicating its direction of change: decreasing, constant, or increasing. *Quantity spaces* define the possible states of the system by determining the range of possible *values* for each quantity, represented as alternating *point* and *interval* values. *Correspondences* (C) can be added to co-occurring values to further determine the possible states of the system. The relationships between quantities are described by two types of causal relationships:

influence and *proportionality*. A causal relationship is of type influence (I) when an active process, indicated by a quantity, is the primary cause of a change in another quantity. This relationship can be either positive (I+) or negative (I−), depending on the directionality of the effect initiated by the process. When the relationship is of type positive, a positive value of the process results in an increase of the related quantity, while a negative value results in a decrease. In cases of a negative influence, a positive value of the process causes a decrease in the related quantity, and a negative value causes an increase. Causal relationships of type proportionality (P) describe how changes in one quantity lead to corresponding changes in another quantity, either in the same direction (P+) or in opposite directions (P−). Exogenous influences are external factors that have a continuous effect on the change of a quantity. In the present paper we restrict to exogenous influences that are either decreasing, constant, or increasing. The behaviour of the system can be further described by (in)equalities, which set ordinal relationships between quantities ($<$, \leq , $=$, \geq , $>$). Calculi allow the execution of qualitative operations such as addition and subtraction.

Simulation within Dynalearn starts with a scenario: the initial settings that define the starting conditions of the model. From these settings, a state graph is generated, visually representing the possible states and transitions of the system. Learners can use this graph to explore and understand the behaviour of the system by navigating through different states. Simulation preferences can be adjusted so that the underlying Garp 3 reasoning engine [11] accounts for possible changes in the first derivative of a quantity (i.e., the second derivative), potentially leading to new states or transitions. Value and inequality history offer an overview of the changes, values and (in)equality of quantities throughout the simulation.

3 Quantitative modelling

In the case of quantitative modelling in secondary education, mathematical equations are used to describe and analyse how systems evolve over time. These models typically use differential equations, linear equations, and nonlinear equations to describe system dynamics.

The differential equation $y(t + \Delta t) = y(t) + m \cdot x(t) \cdot \Delta t$ describes how the value of a function y at time $t + \Delta t$ is derived from its value at a previous time t by adding an increment that depends on the constant m , the value of $x(t)$, and the time step Δt . This formulation uses Euler's method, a finite difference approach commonly used in simulations to approximate the solutions of differential equations. Note, that $x(t)$ itself is a function of time, and its behavior directly influences the behavior of $y(t)$. For example, if $x(t) > 0$ and constant, then $y(t)$ increases linearly. Conversely, if $x(t)$ increases linearly (e.g., $x(t) = mt$), then $y(t)$ exhibits quadratic growth as each increment added to $y(t)$ increases over time. We use Euler here for keeping things simple, though other numerical methods like the Runge-Kutta 4 (RK4) are also commonly employed for more accuracy and

stability. Numerical analysis for solving differential equations is crucial when analytical solutions are not feasible.

Relationships between quantities in a dynamic system can often be described using linear equations, such as $y(t) = m \cdot x(t) \pm b$, where m represents the slope and b is the y -intercept, indicating the value of y when $x = 0$. Here, $x(t)$ denotes the value of x at time t . It's important to note that $y(t)$ exhibits linear behavior relative to $x(t)$; however, the overall behavior of $y(t)$ in terms of time depends on the behavior of $x(t)$. Specifically, $y(t)$ will only show constant behaviour if $x'(t) = 0$ (i.e., if $x(t)$ is constant over time). For example, in modelling a dynamic system that describes the behaviour of gases, the relationship between temperature and pressure is typically linear under constant conditions.

Non-linear equations describe scenarios in dynamic systems where quantities appear as exponents, products, or other non-linear combinations. For example, the non-linear equation $y(t) = -m \cdot x(t)^2 + b$ illustrates how the intensity of light, $y(t)$, diminishes with the square of the distance, $x(t)$, from a point source as an object moves away over time.

After defining the equations of the dynamic system, a simulation can be initiated. Initial values for the variables must be set, along with the duration of the simulation and the size of each time step. The values of each quantity are then calculated for each time step using an integration method, such as Euler's method.

4 Qualitative vocabulary and mathematical equations

In this chapter, we describe how ingredient types of the qualitative vocabulary relate to mathematical equations. For clarity, when referring to quantities in qualitative representations, we use x, y, z without the time notation t and use the δ symbol to indicate their direction of change. When discussing mathematical equations, we denote these quantities as $x(t), y(t), z(t)$ to specify that they are functions of time, and we use the prime notation to discuss the direction of change of these quantities, for example, $x'(t)$. For discussing time steps, we use the notation Δt , and m and b are used in equations to denote the slope and intercept, respectively.

4.1 Exogenous influence, change, and quantity space

Fig. 1 presents a qualitative representation of quantity x with quantity space $\{0, +\}$ and an increasing exogenous influence acting on it. The initial value of x is zero (0). The simulation result shows two consecutive states: in the first state, x is zero and increasing ($\delta x > 0$), and in the second state (shown), x is positive (+) and continues to increase. The mathematical equation corresponding to the value of x is $x(t + \Delta t) = x(t) + x'(t) \cdot \Delta t$. The quantity space of x defines the range as $x(t) \geq 0$. Given that δx is increasing linearly, $x'(t) > 0$ and remains constant. Conversely, for a constant exogenous influence, $x'(t) = 0$ and remains constant, while for a decreasing exogenous influence, $x'(t) < 0$ and remains constant.

Hence, to transition from a qualitative model to a quantitative one, if the exogenous influence on a variable is increasing or decreasing, then the numerical value of $x'(t)$ must be provided. Additionally, if the initial setting of x starts at an interval, then the initial numerical value of $x(t)$, namely $x(0)$, must also be specified. Furthermore, the value of Δt also needs to be set.

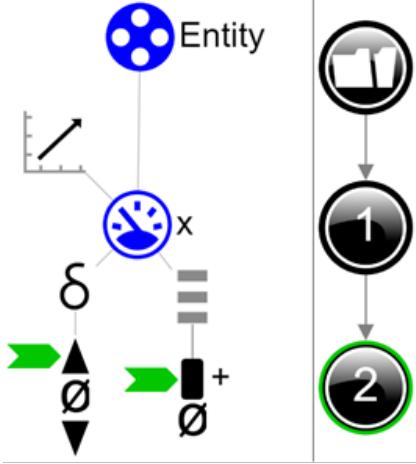


Fig. 1. An increasing exogenous influence acting on quantity x with quantity space $\{0, +\}$. The right side shows the state-graph starting with the scenario followed by two consecutive states. The left side shows the model and the simulation result of the 2nd state (in green).

4.2 Causal relationships

Fig. 2 shows a qualitative representation with a positive proportional relationship (P^+) between quantities x and y , with an increasing exogenous influence acting on x . The simulation result demonstrates that as x increases, y also increases.

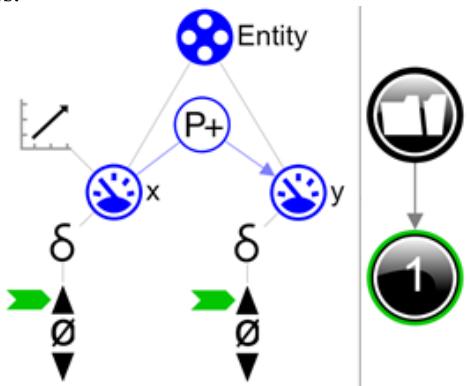


Fig. 2. Positive proportional relationship between x and y .

Assuming a linear relationship between x and y , the general mathematical expression corresponding to this is $y(t) = m \cdot x(t) + b$, and the derivative is $y'(t) = m \cdot x'(t)$. Given the positive proportional relationship, the value of m must be greater than 0. Conversely, for a negative proportional relationship holds $m < 0$. The value of b can be any real number ($b \in \mathbb{R}$), as there are no quantity spaces defined for x

and y that dictate how the values of x and y are related. For further discussion on the latter, see paragraph 4.3.

Fig. 2 could also depict a non-linear positive proportional relationship between x and y , for example dose-response relationship of a certain drug (Fig. 3).

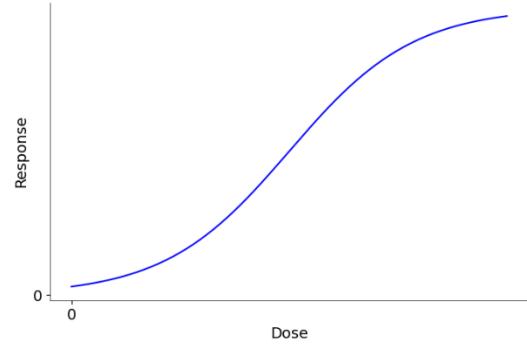


Fig. 3. Non-linear positive proportional relationship between dose and response.

When accounting for non-linear positive proportional relationships, the qualitative representation in Fig. 2 could be described by any mathematical equation whose first derivative is always greater than zero. For example, consider the equation $y(t) = x(t)^3 + 3x(t)$. Following the chain rule, the derivative is $y'(t) = 3x(t)^2 \cdot x'(t) + 3x'(t)$. If $x'(t) > 0$ and remains constant, then $y'(t) > 0$, which indicates that $y(t)$ is a strictly increasing function of $x(t)$.

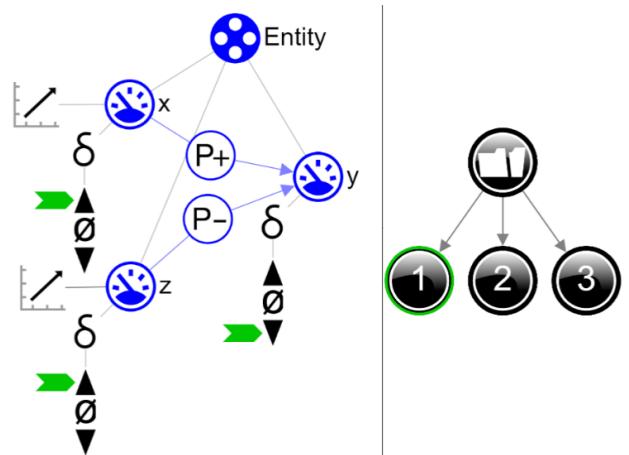


Fig. 4. Positive and negative proportional relationship. The left side shows the model and the simulation result of state 1 (in green).

Fig. 4 shows a qualitative representation where x has a positive proportional relationship with y , and z has a negative proportional relationship with y ; both x and z are increasing due to an increasing exogenous influence. When simulation preferences are set to only consider first changes in the first derivative, the simulation result is ambiguous with three possible final states. In state 1, y is decreasing ($\delta y < 0$); in state 2 (not shown), y is constant; and in state 3 (not shown), y is increasing. The general mathematical equation describing

the change in y , considering a linear relationship between z , x , and y , that corresponds to this representation is $y'(t) = m_1 \cdot x'(t) - m_2 \cdot z'(t)$. Note the minus sign indicates that $z(t)$ has a negative proportional relationship with $y(t)$. The ambiguous simulation result arises not only because m_1 and m_2 are unknown but also because $x'(t)$ and $z'(t)$ are not specified. For example, if $y'(t) = 3x'(t) - 4z'(t)$, and $x'(t)$ is less than $4/3$ times $z'(t)$, then $y(t)$ is decreasing ($y'(t) < 0$). However, if $x'(t)$ is equal or larger than $4/3$ times $z'(t)$, then $y(t)$ is constant or increases. Table 1 shows numerical examples over a time step that illustrate the impact of different ratios of $x'(t)$ and $z'(t)$ on $y'(t) = 3x'(t) - 4z'(t)$. The table demonstrates that if the ratio between $x'(t)$ and $z'(t)$ is 1, then $y'(t) < 0$; if the ratio is $4/3$, then $y'(t) = 0$; and if the ratio is 2, then $y'(t) > 0$.

Table 1. The impact of different ratios of $x'(t)$ and $z'(t)$ on $y'(t)$.

$y'(t) = 3x'(t) - 4z'(t)$									
t	$x'(t) = 1$		$x'(t) = 1$		$x'(t) = 2$				
	$x'(t)$	$z'(t)$	$y'(t)$	$x'(t)$	$z'(t)$	$y'(t)$	$x'(t)$	$z'(t)$	$y'(t)$
0	1	1	-1	1	$-\frac{3}{4}$	0	2	1	2
1	1	1	-1	1	$-\frac{3}{4}$	0	2	1	2
...

Fig. 5 shows the simulation result corresponding to the qualitative representation in Fig. 4, with adjustments in the simulation settings¹ to account for changes in the second-order derivative. These adjustments reveal that transitions between states 1, 2, and 3 are now feasible. Specifically, if one or both relationships of y with x and z are non-linear, the combined effect on δy may depend on specific values. For example, consider if the mathematical equation associated with the qualitative representation of Fig. 4 is $y(t) = x(t)^3 + 3x(t) - 10z(t)$. If both $x(t)$ and $z(t)$ increase consistently (with $x'(t) = 1$ and $z'(t) = 1$) from -3 to 3 , $y(t)$ initially increases, becomes constant, decreases, becomes constant again, and finally increases (Fig. 6). This pattern corresponds to the transitions along path $3 \rightarrow 2 \rightarrow 1 \rightarrow 2 \rightarrow 3$ as shown in the simulation result of Fig. 5.

Fig. 7 shows a qualitative representation of a causal relationship with a positive influence ($I+$) between x and y , with x having quantity space $\{0, +\}$. The simulation result indicates that x is positive and remains constant, which leads to an increase in y ($\delta y > 0$). Note that y does not have a quantity space. Assuming y increases linearly, the corresponding mathematical equation that represents this qualitative relationship is $y'(t) = m \cdot x(t)$. Given the positive influence of x on y , $m > 0$. Furthermore, $x(t) > 0$ and is constant.

Fig. 8 extends the qualitative representation shown in Fig. 7 by including quantity z with a negative influence ($I-$) on y , and now y has quantity space $\{-, 0, +\}$.

¹ We differentiate between initial and simulation settings. The former refers to starting values (and inequalities) when starting a

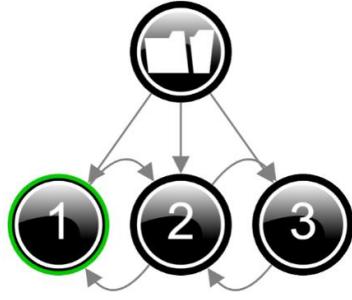


Fig. 5. Ambiguous simulation result with transitions between states 1, 2, and 3.

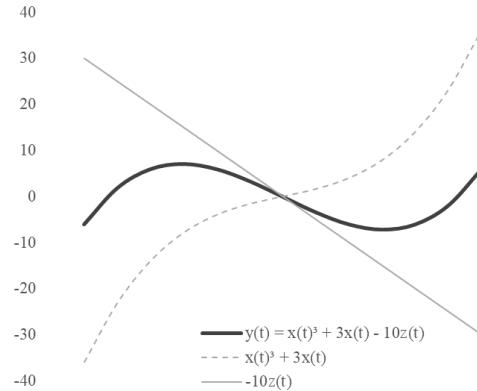


Fig 6. The combined effect of a nonlinear and linear relationship.

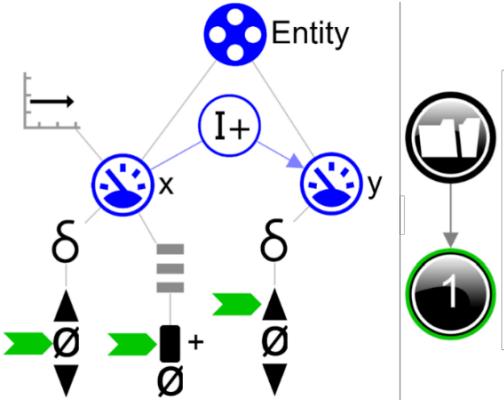


Fig 7. Causal relationship of type influence.

The initial settings are such that y is 0, while both x and z are positive (+) and constant ($\delta x = 0$ and $\delta z = 0$). These settings introduce ambiguity in the simulation result due to the opposing influences: x has a positive effect on y , while z has a negative effect, and their relative magnitudes are unknown. If the influence of z on y is greater than that of x , y will decrease and become negative (path 1 \rightarrow 5); if the influences are equal, y remains at zero (state 2); and if the influence of x is greater than z , y will increase and become positive (path 3 \rightarrow 4).

simulation. The latter refers to characteristics of the reasoning engine [12].

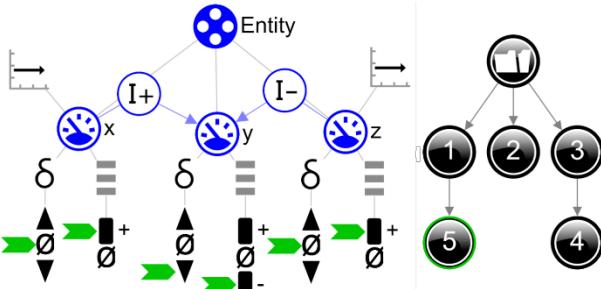


Fig. 8. A negative and positive influence acting on y.

Assuming y increases or decreases linearly, the corresponding mathematical equations are $y(t + \Delta t) = y(t) + m_1 \cdot x(t) \cdot \Delta t$ for the influence of x on y , and $y(t + \Delta t) = y(t) - m_2 \cdot z(t) \cdot \Delta t$ for z 's influence, where $m_2 < 0$ indicating a negative influence. Combining these, the overall expression for $y(t)$ becomes $y(t + \Delta t) = y(t) + (m_1 \cdot x(t) - m_2 \cdot z(t)) \cdot \Delta t$. Here, if $m_1 \cdot x(t) < m_2 \cdot z(t)$, then $y(t)$ decreases; if $m_1 \cdot x(t) = m_2 \cdot z(t)$, then $y(t)$ remains steady; and if $m_1 \cdot x(t) > m_2 \cdot z(t)$, then $y(t)$ increases.

Hence, to transition from a qualitative to a quantitative model, the mathematical equations that describe the causal relationships must be specified. Additionally, the numerical values for the parameters of these equations, such as m and b , must also be provided.

4.3 Correspondence and quantity space

Fig. 9 shows a positive proportional relationship between x and y . An increasing exogenous influence is acting on x , and x has quantity space $\{0, +\}$. Because the quantity space of x includes no negative numbers, any equation for which $y(t)$ is increasing within $x(t) \geq 0$ is valid.

For example, if we assume a linear relationship between x and y , then the general mathematical equation $y(t) = mx(t) \pm b$, with $x(t) \geq 0$ and $m > 0$, is valid. If we assume a non-linear relationship, then $y(t) = x(t)^2$, is also valid. Fig. 10 shows that these two equations are strictly increasing in the range $x(t) \geq 0$. Note that $y(t) = x(t)^2$ would not be strictly increasing if the quantity space included negative values for x .

Fig. 11 extends the representation shown in Fig. 9, now defining quantity spaces $\{0, +\}$ for both x and y . This additional specification for y narrows the scope of the proportional relationship between x and y . The initial values are set with x at zero (0) and y positive (+). The simulation result depicts two consecutive states: In state 1, x is zero and increasing, while y is positive and also increasing. In state 2 (not shown), both x and y are positive and continue to increase.

These initial settings inform the mathematical relationship between $x(t)$ and $y(t)$. Given that at $x(0) = 0$, $y(0) > 0$, assuming a linear relationship, the general mathematical equation would be $y(t) = m \cdot x(t) + b$, where $x(t) \geq 0$ and $y(t) > 0$. Conversely, if the initial values were $x(0) > 0$ and $y(0) = 0$, then the equation would be $y(t) = m \cdot x(t) - b$, with $x(t) > 0$ and $y(t) \geq 0$. If the initial values were $x(0) = 0$ and $y(0) = 0$, then $y(t)$ simplifies to $y(t) = m \cdot x(t)$. Fig. 12

displays line graphs illustrating these three mathematical relationships.

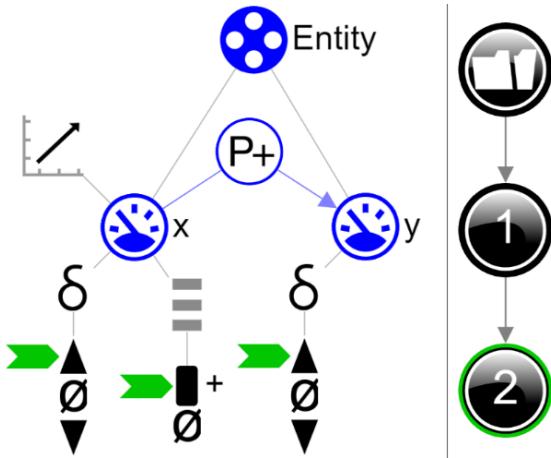


Fig. 9. A positive proportional relationship between x and y , where x has quantity space $\{0, +\}$.

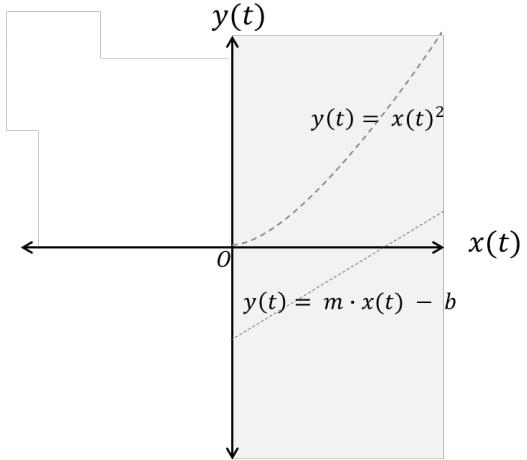


Fig. 10. Examples of linear and non-linear relationships between $x(t)$ and $y(t)$ in the range $x(t) \geq 0$.

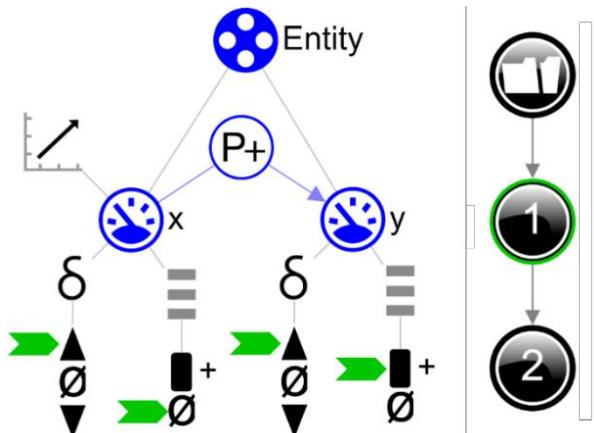


Fig. 11. Both x and y have quantity space $\{0, +\}$. The left side shows the model and the simulation result of state 1 (in green).

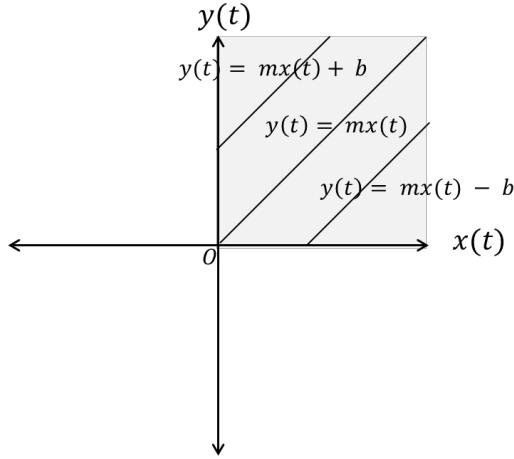


Fig. 12. Three mathematical equations corresponding to different initial settings.

Fig. 13 shows a qualitative representation where both x and y have quantity spaces $\{-, 0, +\}$ and there is a bi-directional correspondence (C) between these quantity spaces. This correspondence defines that if $x = -$ then $y = -$, if $x = 0$ then $y = 0$, and if $x = +$ then $y = +$.

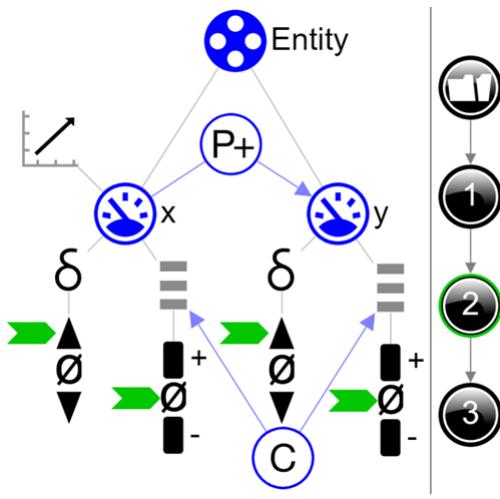


Fig. 13. Bi-directional correspondence between quantity spaces.

Mathematically any equation that goes through the origin and is strictly increasing is valid. For example, if we assume a linear relationship between x and y , then the mathematical equation $y(t) = 2x(t)$ holds. If we assume a non-linear relationship, then $y(t) = x(t)^3 + 3x(t)$ is also valid. Fig. 14 shows that these two equations are strictly increasing.

4.4 Inequality and calculus

Fig. 15 shows a qualitative representation with quantity x with quantity space $\{0, +, \text{transition}, ++\}$ and quantity z with quantity space $\{0, \text{low}, \text{mid}, \text{high}\}$. Quantity x has a positive influence ($I+$) on y and quantity z has a negative influence on y . The initial value for x is ' $++$ ' and for y the initial value is

'low'. There is an (in)equality ($=$) between the 'transition' point from quantity x and 'mid' from quantity z . The (in)equality provides information about the relative size of the influences on y . Given that the value ' $++$ ' for quantity x is above 'transition', and the value 'low' for quantity z is below 'mid', the impact of x on y is greater than that of z . Consequently, the simulation result indicates that y will increase. The corresponding mathematical equation is $y(t + \Delta t) = y(t) + (m_1 \cdot x(t) - m_2 \cdot z(t)) \cdot \Delta t$, with $m_1 \cdot x(t) > m_2 \cdot z(t)$, as also discussed in the accompanying text of Figure 8.

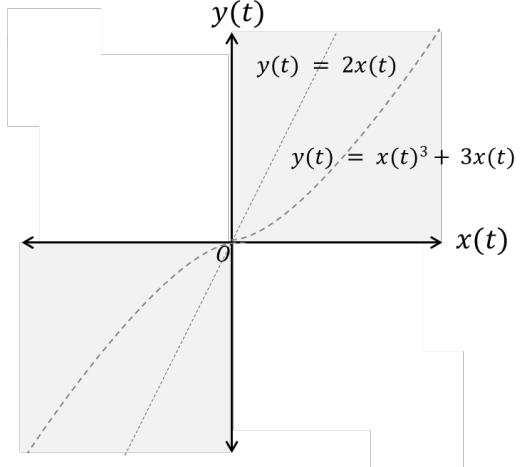


Fig. 14. Two strictly increasing equations that go through the origin.

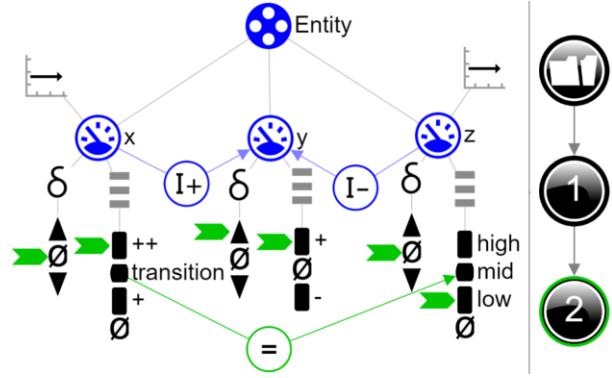


Fig. 15. An (in)equality between two points.

Fig. 16 shows a qualitative representation where x has a positive proportional relationship with y and z has a negative proportional relationship with y . Quantity x and z have quantity space $\{0, +\}$ and y has quantity space $\{-, 0, +\}$. Quantity x has a decreasing exogenous influence acting on it, whereas z has a constant exogenous influence acting on it. There is a calculus that determines that the value of y is the value of x minus the value of z ($y = x - z$). Initially, both x and y are positive (+), with x being greater than y as indicated in the inequality history. The simulation result shows 4 consecutive states. In state 1, x is positive and decreasing, while $x > z$, hence y is positive and decreasing. In state 2, x is still positive and decreasing, x is now equal to z ($x = z$).

Consequently, y is zero and decreasing. In state 3, x is positive, but $x < z$, hence y is negative and decreasing. In state 4, x is zero and steady, thereby y is negative and steady.

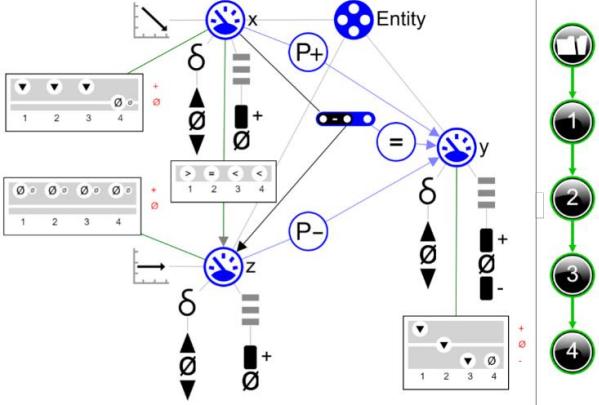


Fig. 16. A calculus specifies that the value of y is x minus z . The grey coloured rectangles show value and inequality histories. For the value history the arrows depict direction of change, the values are shown on the right side, and the state numbers are listed below (e.g., x has value + and is decreasing in state 1). The inequality history depicts the relationship between two quantities (e.g., $x > y$ in state 1, $x = y$ in state 2, etc.).

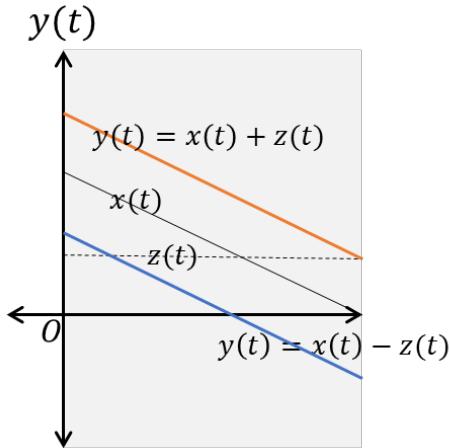


Fig. 17. Lines showing how $y(t)$ changes with $x(t)$ and $z(t)$, both added and subtracted.

The corresponding mathematical equation that models the calculus of the qualitative representation is $y(t) = x(t) - z(t)$, where $x(t) \geq 0$ and $z(t) \geq 0$, and $y(t)$ can be any real number. Note that δx is decreasing linearly, hence $x'(t) < 0$ and remains constant, while $y'(t) = 0$ and remains constant, implying that the rate of change of $y(t)$ is negative ($y'(t) < 0$). Conversely, if the calculus involved addition, as in $y = x + z$, then y would always be positive because z remains positive and x cannot be smaller than zero. Mathematically, if $y(t) = x(t) + z(t)$ and both $x(t)$ and $z(t)$ are non-negative, then $y(t) > 0$. Fig. 17 illustrates the lines corresponding to $x(t)$, $z(t)$, and both $y(t) = x(t) - z(t)$ and $y(t) = x(t) + z(t)$.

5 Dynamics of a falling object as an example

Fig. 18 shows a qualitative representation of the dynamics involved when an object falls and encounters air resistance. The quantities include gravitational force (Fg), air resistance ($Fair$), net force ($Fnet$), acceleration (a), velocity (v), and distance (s), each with a quantity space of $\{0, +\}$. The net force acting on the object is calculated by subtracting air resistance from gravitational force (i.e., $Fnet = Fg - Fair$).

Gravitational force has a positive proportional relationship with net force and air resistance has a negative proportional relationship with net force. Acceleration has a positive proportional relationship with net force, and there is a directed correspondence (C) between the quantity spaces of net force and acceleration. Acceleration has a positive influence on velocity, which in turn positively influences distance. Velocity has a positive proportional relationship with air resistance. The initial settings are that gravitational force has a constant exogenous influence acting in it, velocity and distance are both zero. Acceleration and air resistance derive their values by the directed correspondences.

The simulation of this system with these initial settings shows four consecutive states. In state 1, gravitational force is positive and steady and air resistance is zero and about to increase, resulting in a positive net force ($Fnet > 0$). This positive net force results in acceleration, which in turn causes an increase in velocity ($\delta v > 0$). As the velocity increases, air resistance increases ($\delta Fair > 0$), which decreases the net force ($\delta Fnet < 0$). In state 2, velocity is positive (+) and thereby distance increases ($\delta s > 0$) and air resistance is positive (+). In state 3, distance is positive (+) and increasing ($\delta s > 0$). In state 4, air resistance is equal to gravitational force and the net force is zero ($Fnet = 0$). Thereby acceleration is zero (0) and velocity is positive (+) and constant ($\delta v = 0$).

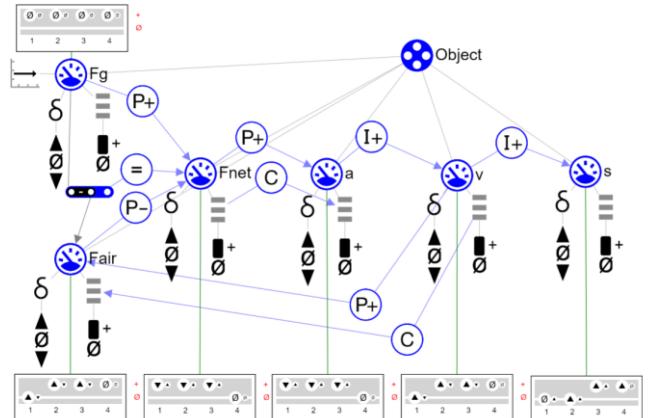


Fig. 18. Qualitative representation of the dynamics involved when an object falls and encounters air resistance. Value history shows the first and second derivative.

To transition from the qualitative representation to an accurate quantitative model, the mathematical equations, along with several initial numerical values for parameters need to be set. The mathematical equations corresponding to the qualitative representation in Fig. 18, which describe the

system of a falling object that encounters air resistance, are detailed in Table 2. The differential equations for velocity $v(t + \Delta t)$ and distance $s(t + \Delta t)$ are linear. The starting values of velocity and distance can be directly taken from the qualitative representation ($v(0) = 0$ and $s(0) = 0$). The equation for gravitational force, $Fg(t + \Delta t)$, is also treated as a differential equation. Typically, in software for numerical simulation, Fg would be considered a constant; however, the vocabulary of Dynalearn does not include an ingredient for constants. The numerical starting value of gravitational force, $Fg(0)$, must be explicitly specified, as it remains constant within a given interval and its exact value is otherwise undefined. In numerical simulations, the parameters mass (m) and the gravitational constant (g) are typically used to calculate the gravitational force acting on the object ($Fg = m \cdot g$). Because Fg is represented in the model as a differential equation but is intended to remain constant, the parameter that governs the increase over time should be set to zero ($c = 0$), ensuring that it does not change. The equation for calculating air resistance incorporates several parameters: Cd is the drag coefficient, which varies based on the object's shape and its movement through the air; ρ represents the air density; and A denotes the cross-sectional area of the object. Additionally, the value of $v(t)$ is squared within this context, reflecting its impact on air resistance as velocity increases.

Table 2. Mathematical equations of the dynamics involved when an object falls and encounters air resistance.

Equations	Initial values
$v(t + \Delta t) = v(t) + a(t) \cdot \Delta t$	$m = .1; g = 9.81$
$s(t + \Delta t) = s(t) + v(t) \cdot \Delta t$	$\rho = 1.3; A = .05; Cd = .3$
$Fg(t + \Delta t) = Fg(t) + c \cdot \Delta t$	$v(0) = 0$
$Fair(t) = \frac{1}{2} \cdot Cd \cdot \rho \cdot A \cdot v(t)^2$	$s(0) = 0$
$Fnet(t) = Fg(t) - Fair(t)$	$Fg(0) = m \cdot g$
$a(t) = Fnet(t) / m$	$\Delta t = .1$
	$c = 0$

Fig. 19 shows the simulation result for velocity per time, based on the equations and initial values listed in Table 2. It shows that velocity starts at zero and increases, aligning with state 1 in Fig. 18. Next, velocity is increasing at a decreasing rate, corresponding to states 2 and 3, before finally stabilizing at a constant value, which corresponds to state 4 in Fig. 18.

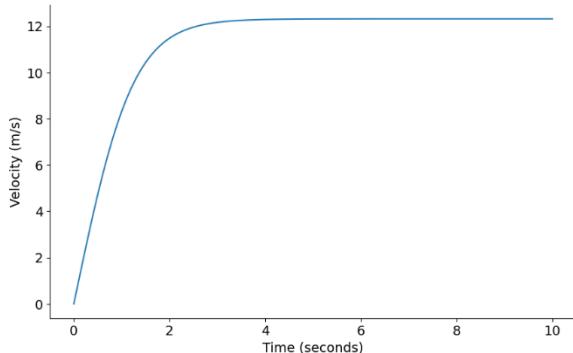


Fig. 19. Simulation result of velocity per time based on equations and initial values of Table 2.

6 Conclusion and future work

In this paper, we focus on how the qualitative vocabulary of Dynalearn, which is used for describing dynamic systems, corresponds to the mathematical equations used in quantitative modeling. We demonstrate how qualitative relationships can be mapped to linear and nonlinear general and differential equations. We also describe how quantity spaces and correspondences define the range of the mathematical equations. The initial values and inequalities set the scenarios in the qualitative representation and provide information about the starting values for parameters in the mathematical equations. Furthermore, a qualitative calculus that specifies operations such as addition or subtraction can be expressed through corresponding mathematical equations.

Dynalearn serves as a learning tool, and for the integration of quantitative modeling, a pedagogical approach should be developed to optimize learning. This approach should include support functions that assist learners in describing the mathematical equations that correspond with the behavior of the qualitative model, as learners often find this challenging [4, 13]. For instance, the software could automatically generate general equations which learners can then edit. For example, the differential equations for $v(t)$ and $s(t)$ as shown in Table 2 could be derived from the quantitative representation in Fig. 18 and presented as the default option.

Another option is to provide feedback based on whether the behavior of the quantitative model aligns with the qualitative model. Since an analytic solution is often not feasible, analysis of whether behaviors align needs to be derived from the simulation result of both models. From the mathematical model, we know that there is no ambiguity in behavior; all values and changes are determined, and the simulation result should at least be a subset of a single path of states from the simulation result of the qualitative model. Remember, a transition in states in the qualitative simulation indicates a change in value or derivative of one or more quantities. To detect changes in the results of the quantitative simulation, it is necessary to check at each time interval whether derivatives change or certain thresholds are reached. If discrepancies are identified between the behaviors, feedback should be provided. For instance, if the results from the quantitative analysis only partially align with a path of states and a final state is not achieved, then the simulation duration may not have been sufficient to reach those subsequent states, or some parameters might need adjustment. For example, if the simulation based on the equations and initial values listed in Table 2 is run for an insufficient duration, the velocity may not stabilize at its final constant state.

With support options in place, the next step is to develop an educational approach that optimizes learning in such integrated software. For instance, a step-by-step approach alternating between qualitative and quantitative modeling, or initially constructing a complete qualitative model to understand system behavior conceptually before transitioning to a quantitative model. Further research on optimizing learning in integrated qualitative and quantitative modeling software is therefore essential.

References

- [1] Jacobson, M., & Wilensky, U. Complex systems in education: Scientific and educational importance and implications for the learning sciences. *The Journal of the Learning Sciences*, 15(1), 11-34, 2006.
- [2] Bredeweg, B., Kragten, M., Holt, J., Kruit, P., van Eijck, T., Pijls, M., Bouwer, A., Sprinkhuizen, M., Jaspar, E., & de Boer, M. Learning with interactive knowledge representations. *Applied Sciences*, 13(9), 5256, 2023.
- [3] Fisher, D. Reflections on teaching system dynamics modeling to secondary school students for over 20 Years. *Systems*, 6(2), 12, 2018.
- [4] Sins, P., Savelsbergh, E., van Joolingen, W. R. The difficult process of scientific modelling: An analysis of novices' reasoning during computer-based modelling. *International Journal of Science Education*, 27(14), 1695-1721, 2005.
- [5] Kragten, M., Spitz, L., & Bredeweg, B. Learning domain knowledge and systems thinking using qualitative representations in secondary education (grades 9-10). In *Proceedings of the 34th International Workshop on Qualitative Reasoning*. Montreal, Canada, 2021.
- [6] Kragten, M., Jaspar, E., & Bredeweg, B. Learning domain knowledge and systems thinking using qualitative representations in secondary education (grades 10-12). In *Proceedings of the 35th International Workshop on Qualitative Reasoning*. Vienna, Austria, 2022.
- [7] Spitz, L., Kragten, M., & Bredeweg, B. Learning domain knowledge and systems thinking using qualitative representations in secondary education (grades 8-9). In *Proceedings of the 34th International Workshop on Qualitative Reasoning*. Montreal, Canada, 2021.
- [8] Forbus, K. D. *Qualitative representations: How people reason and learn about the continuous world*. MIT Press, 2019.
- [9] Coyle, G. Qualitative and quantitative modeling in system dynamics: some research questions. *System Dynamics Review*, 16(3), 225-244, 2000.
- [10] Bredeweg, B., Liem, J., Beek, W., Linnebank, F., Gracia, J., Lozano, E., Wißner, M., Bühlung, R., Salles, P., Noble, R. A., Zitek, A., Borisova, P., & Mioduser, D. (2013). DynaLearn - An intelligent learning environment for learning conceptual knowledge. *AI Magazine*, 34(4), 46-65.
- [11] Bredeweg, B., Linnebank, F., Bouwer, A., & Liem, J. Garp3—Workbench for qualitative modelling and simulation. *Ecological informatics*, 4(5-6), 263-281, 2009.
- [12] Bredeweg, B., & Linnebank, F. E. Simulation preferences - Means towards usable QR engines. In *Proceedings of the 26th International Workshop on Qualitative Reasoning*. Playa del Rey, CA, USA, 2012.
- [13] Buuren, O., Heck, A., & Ellermeijer, T. Understanding of Relation Structures of Graphical Models by Lower Secondary Students. *Research in Science Education*, 46, 633 – 666, 2016.

Reconstructing Qualitative Model Variations from Qualitative Descriptions for Conversational Explanation

Moritz Bayerkuhnlein^{a,b,*} and Diedrich Wolter^b

^aUniversity of Bamberg, Germany

^bUniversity 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

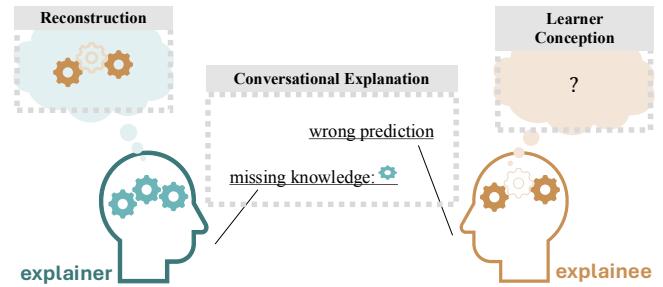


Figure 1. Overview: explainer (teacher) reconstructing the mental model of explainee (learner) before answering with relevant knowledge

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.

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

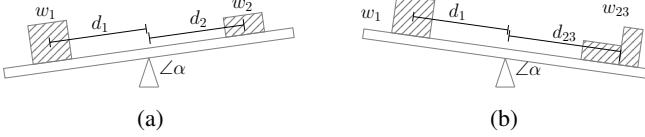


Figure 2. Seesaw configurations

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 $q_1, \dots, q_n \in \mathcal{Q}$.
- \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 , **Proportionality** $P^\pm(q_i, q_j)$ 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 $\text{val}(q, t_i)$ 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 quantity space, 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 sub-graph, 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.

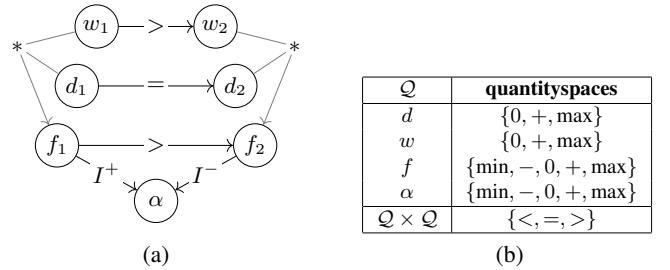


Figure 3. Qualitative Model of a Seesaw depicted in Figure 2a

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 $\text{val}(\alpha_0, t_0) = 0, \delta(\alpha_0, t_0) = 0 \subset s_0 \mid s_0 \in \pi$. 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 $\text{val}(\alpha, t_1) = 0$ to $\text{val}(\alpha, t_2) = +$, 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 $\text{val}(\alpha, t_n) = \max$.

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 *abducting* 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:*

$$S \cup \mathcal{QM} \vdash \text{OBS} \quad (1)$$

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

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

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 quantity 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* $\widehat{\mathcal{QM}}$.

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:

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

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

$$|H \cup R| \text{ is minimal} \quad (5)$$

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 \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.

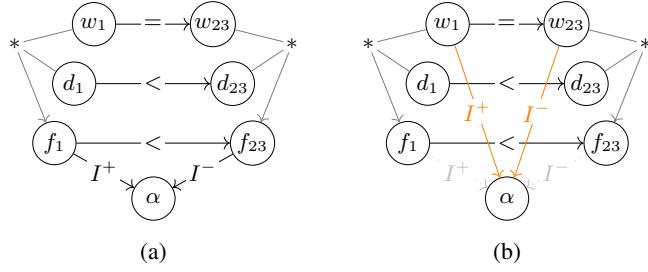


Figure 4. Qualitative Model of a Seesaw depicted in Figure 2a

A learner’s prediction like “the seesaw will be balanced” can then be stated as an observation on their hidden learner model \mathcal{QM} . An observation is a partial state description $\{\text{val}(\alpha, t_n) = 0\}$. This is underconstrained, a more restrictive interpretation of the utterance is $\forall 0 \leq h \leq n: \neg(\text{val}(\alpha, t_h) \neq 0)$, 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:

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

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 $\mathcal{F} = \langle \mathcal{QM}, s_0, S, \mathcal{L}_{QM} \rangle$ where*

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

Definition 4 (Contrastive Explanation Problem). *Given an explanation frame $\mathcal{F} = \langle \mathcal{QM}, s_0, S, \mathcal{L}_{QM} \rangle$, 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 $\widehat{\mathcal{QM}}$, 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.

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

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_ψ 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

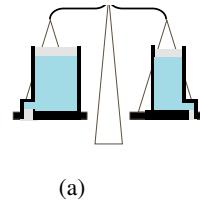
We have implemented qualitative simulations using graph models in Answer Set Programming as a prototype, where the dynamics of the simulations is encoded in rules such as R1 in a logic program. The implementation can generate scenarios, complete partial states to complete states, and generate a full state graph using *brave* enumeration, realizing *prediction*, *postdiction* 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}_{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, \dots, 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 container 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:

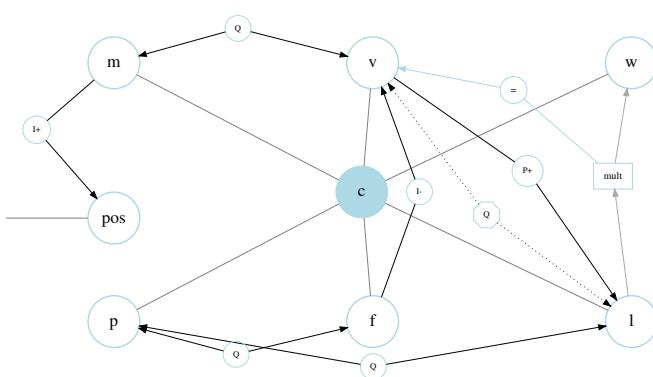


Figure 6. Qualitative Simulation Model excerpt of right container (c) as reference model \mathcal{QM} . Q-nodes denoting correspondences, dotted lines conditionals and white node quantities

$$C_\phi = \{C(l, p), C(p, f), C(l_R, p_R)_{t_2}, C(p_R, f_R)_{t_2}\}$$

$$C_\psi = \{C(v_L, m_L)_{t_2}, I^-(m_L, pos)_{t_1}, I^+(m_R, pos)_{t_1}\}$$

$$C_\Delta = \{C(l, p)\}$$

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 ($C(p, f)$).*

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 explainees' 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 Bundesministerium 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 32nd 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.

Safe parking of a nonholonomic autonomous vehicle by qualitative reasoning

Domen Šoberl^{a, b,*}, Jan Lemeire^{c,d,**}, Ruben Spolmink^d and Jure Žabkar^a

^aUniversity of Ljubljana, Faculty of Computer and Information Science

^bUniversity of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies

^cDept. of Industrial Sciences (INDI), Vrije Universiteit Brussel (VUB), Pleinlaan 2, B-1050 Brussels

^dDept. of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Pleinlaan 2, B-1050 Brussels

Abstract. Traditional techniques for autonomous driving nonholonomic (car-like) vehicles require precise kinematic models and complex geometric computations of trajectories. Learning such a model through reinforcement learning is highly sample inefficient and thus not always feasible in practice. Moreover, such an approach offers poor explainability. We propose an approach based on qualitative reasoning, where a qualitative model for driving a car-like vehicle is learned over a small set of numerical traces. We define a planning algorithm that is able to interpret the learned qualitative models and quantify the actions to pursue the goal while avoiding collisions. We demonstrate our approach on the problem of reverse parallel car parking. The results show that our qualitative approach is able to deduce an S-shaped trajectory to park the car in one smooth reverse maneuver without the typical backward-forward corrections with negligible error in the final position and orientation.

1 Introduction

Nonholonomic vehicles, which include various types of wheeled robots and autonomous vehicles, are subject to constraints that limit their motion to certain paths. Unlike holonomic systems, which can move freely in any direction, nonholonomic vehicles can only move in specific directions due to their constraints. Parking such vehicles involves finding feasible paths that comply with these motion constraints while achieving precise final positioning.

Traditional techniques for autonomous parking of vehicles rely on a combination of sensors (e.g. ultrasonic sensors, cameras, lidar, and radar) and algorithms for path planning and trajectory generation, along with control systems (e.g. PID controllers and Model Predictive Control) to ensure precise vehicle movement and obstacle avoidance. In practice, autonomous parking systems should also take extra care, when dynamicity is present in the environment (e.g. other moving subjects and/or objects nearby). Such methods require a precise kinematic model and are often computationally complex [13]. Recently, Boyali and Thompson [1] proposed a method for optimal path generation in parking maneuvers using a kinematic car model. Their approach integrates Successive Convexification (SCvx) algorithms and state-triggered constraints to ensure path feasibility and constraint satisfaction in constrained environments. Shahi and Lee [14]

introduced a method for autonomous rear parking using Rapidly Exploring Random Trees (RRT) and Model Predictive Control (MPC).

Fundamental geometric methods for generating paths in obstacle-free environments were first studied by Dubins [4]; his paper provides early insights into nonholonomic path planning by studying the shortest paths for car-like vehicles, which can only move forward. Reeds and Shepp [12] addressed also the backward motion of a vehicle. These two studies form the basis for many modern path-planning algorithms used in autonomous vehicles. A basic understanding of motion planning for nonholonomic vehicles is given in Triggs [17].

Alternative approaches use reinforcement learning [19] or fuzzy-based controllers [11] to obtain a good parking strategy, where the vehicle continuously learns from several parking attempts. Reinforcement learning approaches require lots of data and trials, which is not feasible in practice. A recent approach by Moreira [10] explored the application of deep reinforcement learning (DRL) in automated parking. The study focused on training an agent to follow predefined complex paths while avoiding collisions with obstacles.

Commercial autonomous parking systems (APS) can be divided into two types; systems like Bosch's Automated Valet Parking (AVP) also depend on vehicle-to-infrastructure communication to ensure efficiency and safety. For example, Bosch in collaboration with Mercedes-Benz developed an AVP system that enables vehicles to park in predefined parking spots in garages without driver input. These systems exceed the scope of our work. Other brands don't rely on the outside infrastructure: most notably BMW, Audi and Tesla have incorporated APS that use a combination of cameras and ultrasonic sensors to guide the vehicle autonomously. While most of them work well in structured environments like parking garages, real-world scenarios with unpredictable elements (e.g., pedestrians or dynamic obstacles) still present a significant challenge. Vision-based systems, e.g. like those used by Tesla, struggle with low-light conditions, bad weather conditions, and occlusions (e.g., objects blocking sensors). The removal of ultrasonic sensors in some models has also led to inconsistent performance in tight parking spaces.

In this paper, we address the problem of parking a nonholonomic vehicle using qualitative models in combination with qualitative reactive planning [21]. Qualitative models [5, 7, 2, 6] describe the dynamics of a system in qualitative terms such as the directions of change of state variables (increasing, steady or decreasing). These qualitative models can be used in planning and control [15, 9]. We obtain the qualitative model from a small set of numerical traces and

* Corresponding Author. Email: domen.soberl@famnit.upr.si.

** Corresponding Author. Email: jan.lemiere@vub.be.

then use a reactive planning approach to pursue the goal.

This work is part of our endeavor to develop a global learning and planning architecture that can adapt to novel situations in a way that is close to how humans learn. Reverse parallel parking is an interesting challenge to steer this development.

2 The parallel parking challenge

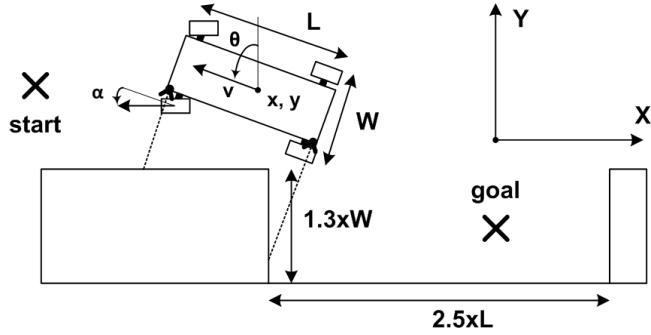


Figure 1. Problem Specification. Here, α is positive and θ is negative. The car width W is $3/5$ of the car length L . The car has 2 distance sensors on its left. The distance they measure is shown with dashed lines.

The experimental setup that we used in this paper is shown in Fig. 1. The car starts at the left of the goal position and drives backward towards the parking spot. The initial and the goal positions are marked in the figure with the cross symbol, which depict the center of the car (x, y). The car's orientation θ at the start is 90 degrees (facing left) and should also be 90 degrees when parked. In this experiment, we constrain the speed of the car v to backward driving at a constant speed ($v < 0$), so the car must be parked in one smooth trajectory, with no back-and-forth maneuvers. This way, the actions are simplified only to turning of the steering wheels within $\alpha \in [-30^\circ, 30^\circ]$. We also place obstacles (other parked cars or walls) in front and behind the parking spot. The obstacle in front is placed at a large enough distance so that the car can perform an S-shaped trajectory in a single maneuver.

There are two distance sensors mounted on the car, one at the front and the other at the back, both on the left side of the car, so that they measure the distance to the nearest wall in the direction perpendicular to the car's orientation (see Fig. 1). The sensor is triggered if the obstacle is closer than the length of the car.

We conducted the experiments in a simulator with a time step of $\Delta t = 40$ milliseconds. The car starts driving backward immediately at a constant speed so that one length of the car is traversed in 20 steps. The speed of turning the steering wheels is 100 degrees per second. Actions therefore only define the direction of turning the steering wheel α , which can be either 0 (no turning), 1 (turning left), or -1 (turning right). The episode stops when the x -position of the car reaches or surpasses the x -position of the parking spot, or when an obstacle is being hit.

3 The numerical model

To simulate the motion of the car, we use a mathematical model similar to the Dubins car model [3], which is often represented as a bicycle model. For an ordinary car, the pairs of parallel wheels are depicted as a single wheel. The car cannot move sideways, and its forward motion is constrained to geometric arcs, as shown in Fig. 2.

The future position of the front wheels is determined by the car's current orientation θ , the distance l between the front and the rear wheel, the current forward velocity v of the car, and the steering angle α , which we constrain to $\alpha \in [-30^\circ, 30^\circ]$. We use the following differential equations to model the dynamics of the front wheel:

$$\begin{aligned}\dot{\theta} &= v \cdot \frac{\sin(\alpha)}{l/2} \\ \dot{x} &= -v \cdot \sin(\theta + \alpha) \\ \dot{y} &= v \cdot \cos(\theta + \alpha)\end{aligned}\quad (1)$$

After the position of the front wheel is calculated for the next time step, the position of the rear wheel is deduced from the new orientation θ and the length l . The midpoint of the segment l is taken as the current car's position (x, y).

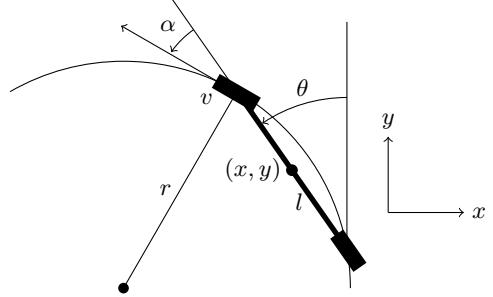


Figure 2. Our mathematical model of a car, which is similar to the Dubins car model.

Differential equations (1) or similar are typically used to model the dynamics of the Dubins car (see, e.g. [8]). However, the Dubins model is constrained to forward motion ($v \geq 0$). The dynamics of moving backward is considerably more complicated. Consider the situation depicted in Fig. 3. The car is oriented towards the left ($\theta = 90^\circ$), drives backwards ($v < 0$), and the steering angle is positive ($\alpha > 0$). The front wheel follows the dynamics of the model (1), which predicts $\dot{y} > 0$, but the rear wheel exhibits the opposite dynamics $\dot{y} < 0$. If such motion is observed long enough, the front wheel will eventually, due to the change in θ , also assume $\dot{y} < 0$. When driving backward, we consider two types of effects: *short-term* effects that describe the immediate dynamics of the front wheel, and *long-term* effects that describe the motion of the back of the car. When parking the car backward, we are interested in the latter.

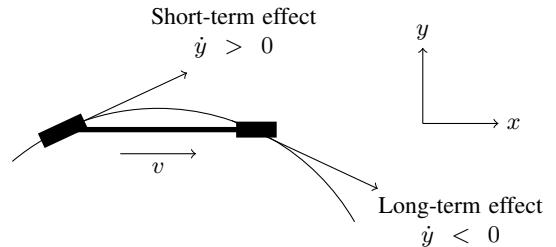


Figure 3. The difference between the *short-term* and the *long-term* action effect.

4 The qualitative model

4.1 A short-term model

The short-term qualitative model can be derived directly from differential equations (1). In this paper we use Q-constraints as defined in

[18]:

$$y = Q^+(x) \quad \text{means} \quad \frac{\partial y}{\partial x} > 0,$$

$$y = Q^-(x) \quad \text{means} \quad \frac{\partial y}{\partial x} < 0.$$

Other functional dependencies may exist, but they are not presumed with the above Q-constraints. We presume that v is constant and $\alpha \in [-30^\circ, 30^\circ]$. Our *short-term* qualitative model for forward driving ($v > 0$) is therefore:

$$\dot{\theta} = Q^+(\alpha),$$

$$\dot{x} = \begin{cases} Q^-(\theta + \alpha) & \text{if } -90^\circ \leq (\theta + \alpha) < 90^\circ, \\ Q^+(\theta + \alpha) & \text{otherwise.} \end{cases}$$

$$\dot{y} = \begin{cases} Q^-(\theta + \alpha) & \text{if } 0^\circ \leq (\theta + \alpha) < 180^\circ, \\ Q^+(\theta + \alpha) & \text{otherwise.} \end{cases}$$
(2)

And for backward driving ($v < 0$):

$$\dot{\theta} = Q^-(\alpha),$$

$$\dot{x} = \begin{cases} Q^+(\theta + \alpha) & \text{if } -90^\circ \leq (\theta + \alpha) < 90^\circ, \\ Q^-(\theta + \alpha) & \text{otherwise.} \end{cases}$$

$$\dot{y} = \begin{cases} Q^+(\theta + \alpha) & \text{if } 0^\circ \leq (\theta + \alpha) < 180^\circ, \\ Q^-(\theta + \alpha) & \text{otherwise.} \end{cases}$$
(3)

The interpretation of the above models is as follows. Consider again the short-term effect in scenario from Fig. 3. The orientation of the car is $\theta = 90^\circ$ and $\alpha \in [-30^\circ, 30^\circ]$. Since $(\theta + \alpha) \in [60^\circ, 120^\circ]$, it applies $\dot{y} = Q^+(\theta + \alpha)$. If we are driving slow, so that v approaches 0, it follows from (1) that $\dot{\theta}$ also approaches 0, hence with slow driving, our Q-constraint approximates $\dot{y} = Q^+(\alpha)$, which we interpret as:

If the steering angle α increases/decreases and everything else remains constant, the speed \dot{y} also increases/decreases.

In our scenario, this means that turning the steering wheel *left* increases \dot{y} , and turning it *right* decreases \dot{y} .

4.2 A long-term model

When driving backward, we use the long-term qualitative model. It is easy to see that short-term and long-term effects on $\dot{\theta}$ are the same, hence $\dot{\theta} = Q^-(\alpha)$ for $v < 0$. However, the long-term effects on \dot{x} and \dot{y} are not directly deducible from the mathematical model (1) without considering some additional geometric properties of the car. We therefore decided to learn the long-term model instead of deducing it. We used the method called Padé [18] that learns Q-constraints from numerical samples.

We collected 330 samples that uniformly cover the domain $\theta \times \alpha$, as seen in Fig. 5. For each configuration (θ, α) , we measured the changes Δx and Δy , while driving backward ($v < 0$) for long enough to observe the long-term effects. Taking into account the duration Δt of each action, we translated the observed values to \dot{x} and \dot{y} . The two outputs from Padé — The first one for \dot{x} and the second one for \dot{y} — are shown in Figure 5. Padé labels each sample with the '+' or the '-' sign, which respectively denote $Q^+(\theta + \alpha)$ and $Q^-(\theta + \alpha)$.

Revisiting again the scenario from Fig. 3, we first identify the qualitative sign belonging to the car's configuration $\theta = 90^\circ$, for some $\alpha > 0$. It is clear from the plots that the long-term effect on \dot{y} of driving backward in this configuration is determined by constraint $\dot{y} = Q^-(\theta + \alpha)$, which means that turning the steering wheel *left*

(increasing α) results in decreasing the speed \dot{y} while turning it *right* (decreasing α) results in increasing the speed \dot{y} .

Fig. 4 gives and interpretation of short-term and long-term qualitative effects on variables y in different states (α, θ) . By turning the wheel, we change the value of alpha either in positive (right arrow) or negative (left arrow) direction. This affects the speed with which the orientation of the car (θ) is changing while driving backward. For a short while (shorter arrow), the sign of y is preserved, but after some time (longer arrow) the sign of y may change.

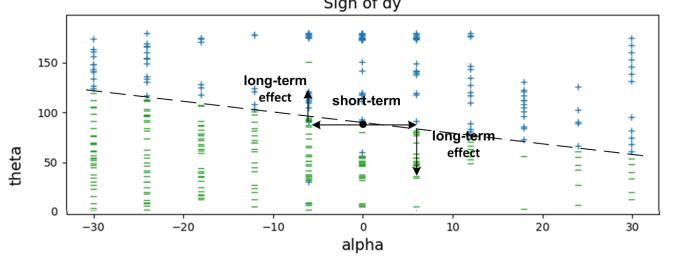


Figure 4. Short-term and long-term action effects observed with the signs of variable y in different states (α, θ) .

5 The planning algorithm

To solve a numerical problem using qualitative models, some form of quantification is necessary. A qualitative model can predict which actions will work in the direction towards a goal state, but cannot assert the quantities of actions or decide on their duration. Using a qualitative model we may decide that the value of some output should be increased or decreased, but cannot directly assert the actual rate of change. In our planning algorithm, we tackle this problem by a reactive approach, where the current numerical state is observed multiple times per second, and each time a qualitative action is decided and executed using a small fixed numerical step. By fast consecutive execution of such short actions, the state of the system is controlled dynamically and steered towards the goal direction. In our car parking domain, the speed of turning the front wheels is fixed, so an action merely decides whether — according to the currently observed state — the driver should be turning the steering wheel left or right.

To decide which action should be executed in some specific moment, we consider the current intention of the driver, which could be one of the following two:

- *The goal pursuit mode.* The collision sensors are off and the goal is to park the car to the designated parking spot.
- *The safety mode.* One or more of the collision sensors got triggered. Avoid colliding with the obstacle/wall.

When avoiding collision, the algorithm temporarily ignores the primary goal of parking the car, until the danger of colliding is over.

5.1 The goal pursuit mode

When pursuing the goal, the planner decides on the next action based on the direction and the distance of the parking spot. There are three spatial variables to consider: x , y , and θ , each with its own goal value. Using a qualitative model, the planner may, for example, deduce that turning the steering wheel left may work in favor of variable y , but unfavorably for variable θ . The priority is then given to variables that

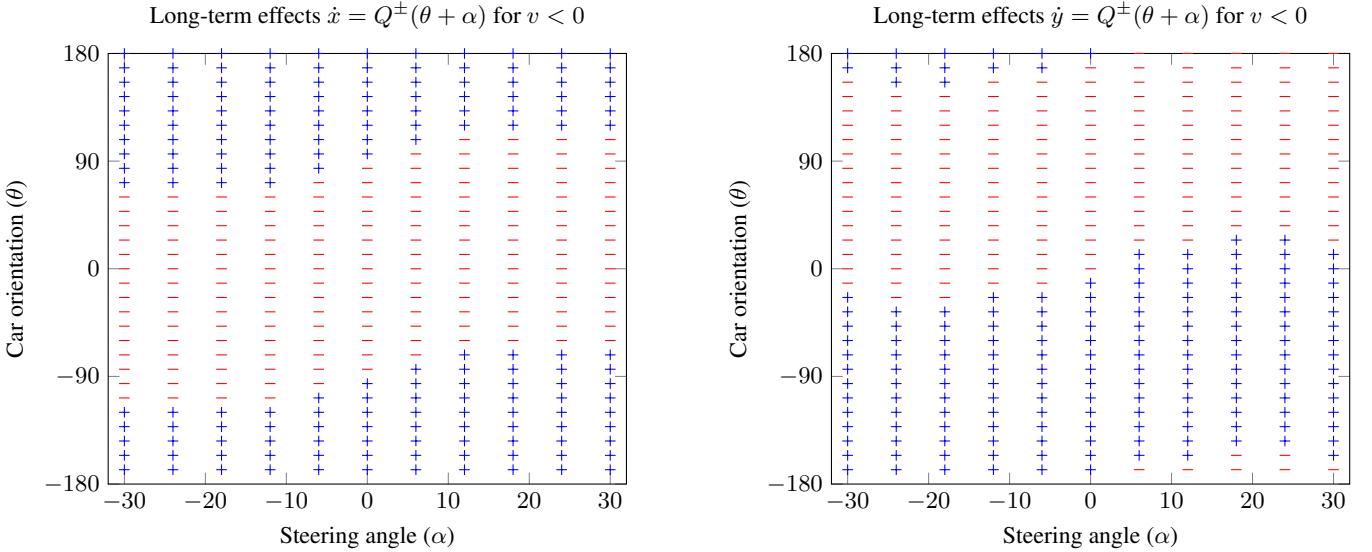


Figure 5. The learned long-term qualitative effects of using different steering angles (α) in different car orientations (θ) while driving backward ($v < 0$). The ‘+’ and the ‘-’ signs respectively denote the Q^+ and the $\dot{y} = Q^-$ constraint.

are farther away from their goal values. This is done in the following way:

1. Observe and store the current speed and acceleration for each variable separately.
2. Using the history of observation, compute theoretically the fastest possible time η_i for each variable x_i to reach its goal value, taking into account the highest observed acceleration $\pm a_{\max}^i$ for each direction separately, the current velocity v_0^i and the observed terminal velocity v_{\max}^i .
3. Use the qualitative model (long-term model for backward driving) to determine how each action would affect the direction of change of each variable.
4. Let each variable x_i use its η_i as a voting score for each action. If the action moves the value x_i towards its goal, the $+\eta_i$ is cast for the action, and if the action moves it away from the goal, $-\eta_i$ is cast.
5. Sum up all the votes for each action and execute the one with the highest score.

Computing the time estimates η_i instead of using actual spatial distances bridges the gap between different units (e.g., meters for x , y , and degrees for θ), while also accounting for different kinematic properties (e.g., rotations could be slow in comparison to forward/backward motion). This way the planner dynamically adapts to the numerical properties of the system. Moreover, by considering this temporal component, the planner aims to bring all variables to their goal values simultaneously. The experiments with a similar approach in [22, 20] show that the action selection algorithm first brings all the η_i values approximately to the same value and then simultaneously lowers them to $\eta_i = 0$ (meaning that the goal state has been reached), if such a behavior is possible. This applies to our parking problem as the capability to park the car in one smooth trajectory without the need for additional corrections, if such a trajectory is possible with the given steering constraints. However, there is no guarantee that the obtained trajectories are optimal.

5.2 The safety mode

When one or more sensors are triggered, the algorithm switches to the *safety mode*, where the goal stops being pursued and the aim is to avoid collision. In some of the previous work (e.g., [21, 22]), collision avoidance has been successfully executed while simultaneously pursuing the goal. However, it was only shown to work with point obstacles and a sensory input that exhibits continuous changes in the input values. In our parking domain, sensory input is typically not continuous — a wall may come to an end, and the input may instantly jump from, e.g., 0.5 meters to infinity. This confuses the η_i computations with erroneous observations of velocities and accelerations, so the sensory variables cannot be compared with the pursuit variables when voting for individual actions. We therefore introduce the *safety mode*, where the actual values of the distance sensors are used instead of η_i , to prioritize actions.

In *safety mode*, actions are decided in the following way:

1. Observe the current values x_i of active sensors (the distance from the obstacle).
2. Use the qualitative model to predict whether an action increases or decreases the sensor’s distance to the obstacle.
3. Vote by $1/x_i$ for an action, if the action increases the distance from the obstacle, and by $-1/x_i$ if it decreases it.
4. Sum up all the votes for each action and execute the one with the highest score.

In our parking domain, we use the short-term qualitative model for the front sensor and the long-term qualitative model for the rear sensor. The reasons are obvious from Fig. 3.

6 Experimental Results

The proposed planning approach succeeded in parking the car without a collision. Fig. 6 shows the result of a simulation at 25 FPS, which took 80 steps (3.2 seconds).

In the beginning, both sensors turn on because of the proximity of the left-side obstacle. The car therefore drives straight back until the rear sensor turns off. Still in safety mode, a slight clockwise turn is made to increase the distance of the front sensor from the obstacle, and soon after the front sensor also turns off. The car continues with pursuing the goal and makes an S-curved trajectory towards the goal position. The parking finishes with the goal orientation error of 2.9 degrees (final θ was 87.1°).

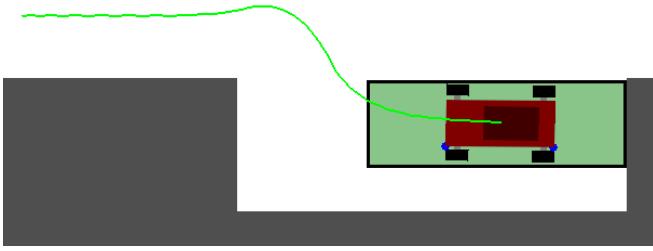


Figure 6. Successful parking maneuver with the proposed approach. The green rectangle denotes where we want the car to be parked. The yellow points mark the points where the distance sensors touch the wall.

Next, we performed 100 experiments to test the efficacy of our method. Initial positions were randomly chosen. The x -position was varied with a maximum deviance of twice the length of the car and chosen so that the back of the car was not past the first corner. The y -position was varied with a maximum deviance of the length of the car with a minimal distance to the wall of $\frac{1}{4}$ of the car's width. The initial orientation θ was chosen within $[60^\circ, 120^\circ]$, thus with a maximal deviation of 30° from a perfect parallel orientation. The results are shown in Fig. 7. The arrows show the initial positions, such as the arrow in Fig. 1. Green arrows indicate a successful parking maneuver, while red indicates failures. 89 out of the 100 experiments were successful.

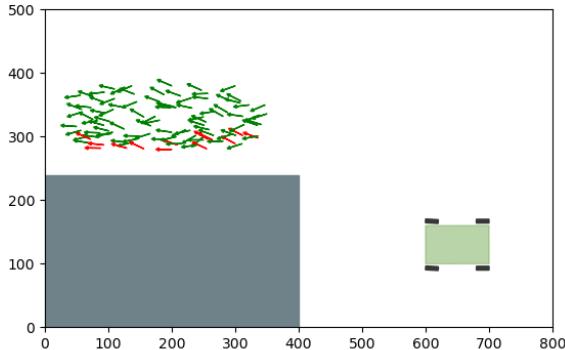


Figure 7. The results of 100 experiments with different initial states. Green and red arrows respectively indicate initial positions and orientations of successful and unsuccessful parking attempts.

There are two patterns of failures. First, if the car is close to the wall and oriented toward the wall with its back side, it cannot make

a smooth backward trajectory without colliding with the wall, due to the constraints on the steering angle (Fig. 8). To resolve the situation, a forward maneuver should be made, which is not allowed by our current experimental setup. The second failing scenario occurs when in safety mode, the car is brought into a position of the first type. In our experimental setup, that would happen after successfully passing the wall on the left and taking a sharp turn left while still in safety mode. When switching back to the goal pursuit mode, the car is positioned too close to the wall to be able to make a smooth trajectory without collision (Fig. 9). In this case also, forward driving to correct the position would resolve the situation.

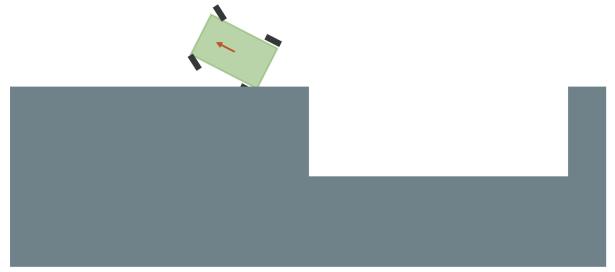


Figure 8. Parking fails if the car is initially too close to the wall and oriented in such a way, that even with the maximum steering angle, collision with the wall is unavoidable.

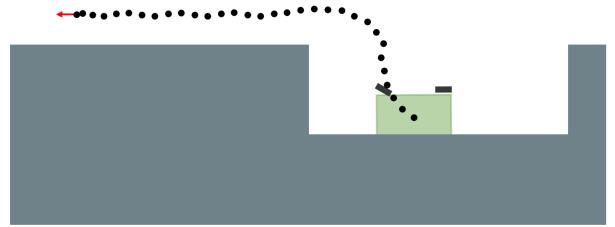


Figure 9. Parking fails, if in safety mode, the car is brought into a position, from which a smooth trajectory is not possible, after switching back to goal pursuit mode.

7 Conclusion

In this paper we showed that the problem of learning how to park a nonholonomic autonomous vehicle can be approached qualitatively. The main advantage of our qualitative approach is significantly higher samples efficiency and the speed of learning a model than with the traditional reinforcement learning methods. We employed a reactive planning method that has already been successfully used with qualitative models for differential drive, quadcopter control, and a cart-pole system [21, 22, 16]. We proposed a novel addition to this type of planning, which is a separation of the *safety model* from the *goal pursuit* model, which solved the problem of discontinuous input from the sensors, as well as previously unaddressed problem of detecting multiple obstacles simultaneously or through

multiple sensors. The results showed the ability of our method to park the car with high accuracy in a single maneuver.

The experiment demonstrated in this paper was simplified by keeping the speed constant at all times, which simplified car actions to only turning the wheel. It would be more realistic to also employ speed regulation with the possibility to also move forwards and stop the car at any time. This would also address the problem of non-determinism of constraints of type $Q^\pm(\theta + \alpha)$, where the actual outcome depends on the rate of change of both, θ and α . By stopping the car, θ can be considered as a constant, hence $Q^\pm(\theta + \alpha)$ becomes equivalent with $Q^\pm(\alpha)$. This also complies with the way humans usually park a car — often stopping the car while turning the steering wheel, so to make for the moment the steering wheel the only deciding factor of the next driving direction. By keeping the speed constant, certain trajectories were not possible that would otherwise be feasible, which also includes collisions that could otherwise be avoided.

Acknowledgements

This work was partially supported by the Slovenian Research Agency (ARRS), grant L2-4436: Deep Reinforcement Learning for optimisation of LV distribution network operation with Integrated Flexibility in real-Time (DRIFT).

References

- [1] A. Boyali and S. Thompson. Autonomous parking by successive convexification and compound state triggers. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1204–1210. IEEE, 2020.
- [2] I. Bratko and D. Suc. Learning qualitative models. *AI Magazine*, 24: 107–119, 01 2004.
- [3] L. E. Dubins. On curves of minimal length with a constraint on average curvature, and with prescribed initial and terminal positions and tangents. *American Journal of Mathematics*, 79(3):497–516, 1957.
- [4] L. E. Dubins. On curves of minimal length with a constraint on average curvature, and with prescribed initial and terminal positions and tangents. *American Journal of Mathematics*, 79(3):497–516, 1957.
- [5] K. D. Forbus. Qualitative process theory. *Artificial intelligence*, 24: 85–168, 1984.
- [6] K. D. Forbus. *Qualitative representations: How People Reason and Learn about the Continuous World*. MIT Press, 2018.
- [7] B. Kuipers. Qualitative simulation. *Artificial Intelligence*, 29(3):289–338, 1986.
- [8] S. M. LaValle. *Planning Algorithms*. Cambridge University Press, USA, 2006. ISBN 0521862051.
- [9] J. Lemeire, N. Wouters, M. Van Cleemput, and A. Heirman. Contextual qualitative deterministic models for self-learning embodied agents. In C. L. Buckley, D. Cialfi, P. Lanillos, M. Ramstead, N. Sajid, H. Shimazaki, T. Verbelen, and M. Wisse, editors, *Active Inference*, pages 3–13, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-47958-8.
- [10] D. Moreira. Deep reinforcement learning for automated parking. *Applied Intelligence*, 51(2):397–416, 2021.
- [11] N. Nakrani and M. Joshi. A human-like decision intelligence for obstacle avoidance in autonomous vehicle parking. *Appl Intell*, 52:3728–3747, 2022.
- [12] J. A. Reeds and L. A. Shepp. Optimal paths for a car that goes both forwards and backwards. *Pacific Journal of Mathematics*, 145(2):367–393, 1990.
- [13] M. Rosenfelder, H. Ebel, J. Krauspenhaar, and P. Eberhard. Model predictive control of non-holonomic vehicles: Beyond differential-drive. *arXiv preprint arXiv:2205.11400*, 2022.
- [14] S. Shahi and H. Lee. Autonomous rear parking via rapidly exploring random-tree-based reinforcement learning. *arXiv preprint arXiv:2205.11400*, 2022.
- [15] D. Šoberl and I. Bratko. Unified approach to qualitative motion planning in dynamic environments. In *Proceedings of the 29th International Workshop on Qualitative Reasoning*, 2016.
- [16] D. Šoberl and I. Bratko. Transferring a learned qualitative cart-pole control model to uneven terrains. In A. Bifet, A. C. Lorena, R. P. Ribeiro, J. Gama, and P. H. Abreu, editors, *Discovery Science*, pages 446–459, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-45275-8.
- [17] B. Triggs. Motion planning for nonholonomic vehicles: An introduction. In *Proceedings of the 1993 IEEE International Conference on Robotics and Automation*, pages 2605–2610. IEEE, 1993.
- [18] J. Žabkar, I. Bratko, and J. Demšar. Learning qualitative models through partial derivatives by Padé. In *Proceedings of the 21st Annual Workshop on Qualitative Reasoning*, pages 193–202, 2007.
- [19] P. Zhang, L. Xiong, Z. Yu, P. Fang, S. Yan, J. Yao, and Y. Zhou. Reinforcement learning-based end-to-end parking for automatic parking system. *Sensors*, 19(18), 2019. ISSN 1424-8220. doi: 10.3390/s19183996. URL <https://www.mdpi.com/1424-8220/19/18/3996>.
- [20] D. Šoberl. *Automated planning with induced qualitative models in dynamic robotic domains*. PhD thesis, University of Ljubljana, 2021.
- [21] D. Šoberl and I. Bratko. Reactive motion planning with qualitative constraints. In *Advances in Artificial Intelligence: From Theory to Practice*, pages 41–50. Springer International Publishing, 2017. doi: 10.1007/978-3-319-60042-0_5.
- [22] D. Šoberl and I. Bratko. Learning to control a quadcopter qualitatively. *Journal of Intelligent & Robotic Systems*, 2020. doi: 10.1007/s10846-020-01228-7.

Using qualitative reasoning to compare media coverage of Israel-Gaza war

Walaa Abuasaker^a, Núria Agell^b, Jennifer Nguyen^b, Nil Agell^b, Jordi Nin^b, Mónica Sánchez^a and Francisco J. Ruiz^a

^aUPC-BarcelonaTech

^bEsade Business School, URL

ORCID (Walaa Abuasaker): <https://orcid.org/0009-0000-4630-6687>, ORCID (Núria Agell):

<https://orcid.org/0000-0001-9264-2147>, ORCID (Jennifer Nguyen): <https://orcid.org/0000-0002-7498-7536>,

ORCID (Nil Agell): <https://orcid.org/0009-0005-0756-5631>, ORCID (Jordi Nin):

<https://orcid.org/0000-0002-9659-2762>, ORCID (Mónica Sánchez): <https://orcid.org/0000-0002-0201-345X>,

ORCID (Francisco J. Ruiz): <https://orcid.org/0000-0002-4314-101X>

Abstract. In this study, we present a methodology based on absolute orders-of-magnitude qualitative reasoning to aggregate and compare news from diverse sources. Our approach integrates linguistic scales to enhance the comprehension of different perceptions and attitudes. We conduct a comparative analysis of news coverage across European countries with respect to the Israel-Gaza war, aiming to capture the sensitivity towards this ongoing conflict.

1 Introduction

News is core information to manage and influence citizens' opinions and interests. In this direction, the news that appears in newspapers on various conflicts and crisis around the world, model citizens' knowledge and perspectives. However, concerning the same event, different opinions are expressed in different contexts. In this article, the information in terms of sentiment is analysed and compared from the perspective of different European countries.

We consider a large-scale news media coverage data set collected by the GDELT (Global Data on Events, Location, and Tone) Project. GDELT covers news media in over 100 languages from the whole world [5]. The data is open-source and is designed to provide a means of analyzing trends and better understand the behaviours behind different types of events. The events are collected from major international, national, regional, and local news sources. Local services and global news agencies also contribute to this platform.

In this paper we represent sentiment tone in a linguistic ordinal scale with unbalanced terms to capture the interest and emotions on the on-line news with respect to the Israel-Gaza war. Traditional qualitative reasoning models address diverse perspectives, such as individual opinion representation and the qualitative fusion of opinions to capture group consensus [4, 11]. These models have advanced within linguistic computation and are particularly applied in contexts where understanding people's emotions or sentiments is of interest [2, 3, 12].

We study and compare news perception among four European countries, specifically, the UK, Germany, France and Spain, towards the Israel-Gaza war during the first 143 days of the war, from 7th October, 2023 to 26th February, 2024. In particular, in this preliminary study, we have considered the following newspapers: Bild (bild.de), Süddeutsche Zeitung (sueddeutsche.de), Die Welt (welt.de), Frankfurter Allgemeine Zeitung (faz.net) and Die Tageszeitung (taz.de)

from Germany; La Croix (la-croix.com), Le Monde (lemonde.fr), Les Echos (lesechos.fr), Libération (liberation.fr), l'Humanité (humanite.fr) and Le Figaro (lefigaro.fr) from France; ABC (abc.es), El Periódico (elperiodico.com), La Razón (larazon.es), El País (elpais.com) and La Vanguardia (lavanguardia.com) from Spain; Daily Mail (dailymail.co.uk), Independent (independent.co.uk), The Guardian (theguardian.com), The Telegraph (telegraph.co.uk) and BBC (bbc.co.uk) from UK.

We consider that the linguistic scales used have a different meaning for each country, and even so, we will be able to aggregate all this information to obtain a global sentiment value and a real comparison between the different sensitivities of countries regarding the war between Israel and Gaza, beyond appearances, due to the different traditions in narrative and the linguistic expressions unique to each language ([7], [8]).

An automated methodology which aggregates and compares news sentiment across countries' is defined based on sentiment analysis and perceptual maps.

The remainder of the paper is organized as follows. Section 2 contains the basic concepts of linguistic perceptual maps, centroid and degree of consensus. Then, in Section 3, we delve into the real case study, analyzing news perceptions of the Israel-Gaza war. We introduce our methodology and discuss the results we obtained. Finally, conclusions, challenges and future research directions are drawn in Section 4.

2 Preliminaries

This section contains the definitions of some preliminary concepts on linguistic perceptual maps based on [1] that are necessary for the methodology presented.

Let S be a totally ordered set of *basic linguistic terms* (BLTs), $S=\{s_1, \dots, s_n\}$, with $s_1 < \dots < s_n$ and we consider the concept of hesitant linguistic terms, which encompasses the intervals of consecutive BLTs.

Definition 1 A *hesitant linguistic term* (HLT) over S is a subset of consecutive BLTs of S , i.e., $\{x \in S \mid s_i \leq x \leq s_j\}$, for some $i, j \in \{1, \dots, n\}$ with $i \leq j$. For completeness, the empty set $\{\} = \emptyset$ is also considered as a HLT and it is called the *empty HLT*.

The non-empty HLTs $\{x \in S \mid s_i \leq x \leq s_j\}$ are denoted by $[s_i, s_j]$. If $i = j$, $[s_i, s_i]$ is the singleton $\{s_i\}$. The set of all non-empty HLTs over S is denoted by \mathcal{H}_S , that is, $\mathcal{H}_S = \{[s_i, s_j] : i, j \in \{1, \dots, n\}, i \leq j\}$. In this way, the set of all HLTs over S is $\mathcal{H}_S \cup \{\emptyset\}$.

Example 1 Let S be a totally ordered set of basic linguistic terms with granularity $n = 4$, $S = \{s_1, s_2, s_3, s_4\}$ being $s_1 = \text{low}$, $s_2 = \text{medium}$, $s_3 = \text{high}$ and $s_4 = \text{very high}$. Given the negative sentiments corresponding to the news of three different newspapers from the same country in a specific day, $A = \text{considerably high}$, $B = \text{low}$, and $C = \text{not low but not very high}$, their respective HLT can be represented as $H_A = [s_3, s_4]$, $H_B = \{s_1\}$ and $H_C = [s_2, s_3]$.

In \mathcal{H}_S , the *set inclusion* relation (\subseteq) provides a partial order. The connected union of two HLTs is defined as the least element of $\mathcal{H}_S \cup \{\emptyset\}$, based on the subset inclusion relation \subseteq , that contains both HLTs. The connected union together with the intersection provide to the set of HLTs, $\mathcal{H}_S \cup \{\emptyset\}$, a *lattice structure*, as proven in [8].

Unlike quantitative values (numbers), the meaning of linguistic labels is not always the same and depends greatly on the context and, above all, on the user's background [6]. For this reason, the concept of *linguistic perceptual map* was introduced in [10] as a normalized measure in the set of HLTs. Different users may handle the same linguistic labels but different perceptual maps.

Let us consider a normalized measure μ over S , i.e., $\mu : S \rightarrow [0, 1]$ such that $\sum_{i=1}^n \mu(s_i) = 1$. For any $s_i \in S$, we call $\mu(s_i) \equiv \mu_i$ the *width* of the basic label s_i . The following definition extends to \mathcal{H}_S the concept of width.

Definition 2 Given $H = [s_i, s_j] \in \mathcal{H}_S$, then the *width* of H is $\mu([s_i, s_j]) \equiv \sum_{k=i}^j \mu_k$. The pair (\mathcal{H}_S, μ) , that we also denote as $\mathcal{H}_{(S, \mu)}$, is called *linguistic perceptual map*.

Any linguistic perceptual map is uniquely associated with a partition of the interval $[0, 1]$ into n sub-intervals of lengths μ_1, \dots, μ_n and also with a set of landmarks $\lambda_0 = 0, \lambda_1, \dots, \lambda_{n-1}, \lambda_n = 1$. The relationship between the landmarks and the width of the basic linguistic labels is $\lambda_m = \sum_{i=1}^m \mu_i$ and $\mu_m = \lambda_m - \lambda_{m-1}$, for any $m = 2, \dots, n$ and $\mu_1 = \lambda_1$.

To compare linguistic terms expressed in different linguistic perceptual maps, in this paper, following the procedure introduced in [1], we consider the common perceptual map that provides a unified context. Although the common perceptual map usually has a higher granularity, it is the adequate framework to represent, fuse and compare different expressions of the same linguistic terms.

Definition 3 Let $\mathcal{H}_{(S^m, \mu^m)}, m \in \{1, \dots, k\}$ a set of k linguistic perceptual maps. Let $\{\lambda_0^m = 0, \lambda_1^m, \dots, \lambda_{n_m}^m = 1\}$, for $m \in \{1, \dots, k\}$, the sets of landmarks of the k partitions associated. The *common perceptual map*, $H_{(S^U, \mu^U)}$, is the linguistic perceptual map associated to the partition, P_U , of landmarks $\bigcup_{m=1}^k \bigcup_{i=0}^{n_m} \{\lambda_i^m\}$. The cardinality of this partition satisfies $N \equiv \#P_U \leq \sum_{m=1}^k n_m - 1$.

In addition, based on the linguistic perceptual maps lattice structure, a perceptual-based distance between HLTs is defined. This distance will allow us to introduce the concept of centroid.

Definition 4 Let $\mathcal{H}_{(S, \mu)}$ be a linguistic perceptual map. Given $H_1, H_2 \in \mathcal{H}_{(S, \mu)}$, the *perceptual-based distance* between H_1 and H_2 is defined as:

$$D_\mu(H_1, H_2) = 2 \cdot \mu(H_1 \sqcup H_2) - \mu(H_1) - \mu(H_2) \quad (1)$$

Example 2 Considering the same three newspapers from Example 1, let's assume that in their country $\mu(s_1) = 0.22, \mu(s_2) = 0.24, \mu(s_3) = 0.26, \mu(s_4) = 0.28$. According to Equation (1), the distances between A, B and C are $D_\mu(H_A, H_B) = 2 \cdot \mu(H_A \sqcup H_B) - \mu(H_A) - \mu(H_B) = 2 \cdot 1 - 0.54 - 0.22 = 1.24$, $D_\mu(H_A, H_C) = 2 \cdot \mu(H_A \sqcup H_C) - \mu(H_A) - \mu(H_C) = 2 \cdot 0.78 - 0.54 - 0.5 = 0.52$ and $D_\mu(H_B, H_C) = 2 \cdot \mu(H_B \sqcup H_C) - \mu(H_B) - \mu(H_C) = 2 \cdot 0.72 - 0.22 - 0.5 = 0.72$.

In [10] it is proved that this definition is indeed a distance in \mathcal{H}_S . The centroid of a set of HLTs is introduced in order to obtain a collective opinion.

Definition 5 Let $\mathcal{H}_{(S, \mu)}$ be a linguistic perceptual map. Let $\{H_m = [s_{L_m}, s_{R_m}] \in \mathcal{H}_{(S, \mu)} : m \in \{1, \dots, k\}\}$ be a set of HLTs, the *centroid of this set*, denoted as H^C , is defined as:

$$H^C = \arg \min_{H \in \mathcal{H}_{(S, \mu)}} \sum_{m=1}^k D_\mu(H, H_m). \quad (2)$$

In [10], it was proved that the centroid can be any term from the set:

$$\{[s_L, s_R] \in \mathcal{H}_{(S, \mu)} \mid L \in \mathbb{M}(L_1, \dots, L_k)\}, \quad (3)$$

$$R \in \mathbb{M}(R_1, \dots, R_k), L \leq R\}$$

where $\mathbb{M}()$ is the set that contains just the median if k is an odd number or the set of two central values and any integer number between them if k is even.

Example 3 Following Examples 1 and 2, the centroid of the three newspapers' negative sentiments is: $H^C = [s_2, s_3]$. Note that, since $H_A = [s_3, s_4], H_B = \{s_1\}$ and $H_C = [s_2, s_3]$, 2 is the median of the set of three left-hand indexes $\{1, 2, 3\}$ and 3 is the median of the set of three right-hand indexes $\{1, 3, 4\}$.

On the other hand, given a set of k linguistic perceptual maps $\{\mathcal{H}_{(S^m, \mu^m)} \mid m \in \{1, \dots, k\}\}$ and its corresponding common perceptual map $\mathcal{H}_{(S^U, \mu^U)}$, the projection of basic linguistic terms is defined in the following way.

Definition 6 Let $S = \{s_1, \dots, s_n\}$ be a set of BLTs and $\mathcal{H}_{(S, \mu)}$ be one of the linguistic perceptual maps from the set $\{\mathcal{H}_{(S^m, \mu^m)} \mid m \in \{1, \dots, k\}\}$ in which, for the sake of simplicity, we avoid the index m . Let $\mathcal{H}_{(S^U, \mu^U)}$ be the common perceptual map with $N = \#S^U$. The *projection function of BLTs* is $\pi : S \rightarrow \mathcal{H}_{(S^U, \mu^U)}$ defined by $\pi(s_i) = [s_{L_i}^U, s_{R_i}^U] \in \mathcal{H}_{(S^U, \mu^U)}$, holding $\sum_{l=1}^{i-1} \mu_l = \sum_{\alpha=1}^{L_i-1} \mu_\alpha^U$ and $\sum_{l=i+1}^n \mu_l = \sum_{\alpha=R_i+1}^N \mu_\alpha^U$, for each $i \in \{1, \dots, n\}$.

Definition 7 Let $S = \{s_1, \dots, s_n\}$ be a set of BLTs and $\mathcal{H}_{(S, \mu)}$ be one of the linguistic perceptual maps from the set $\{\mathcal{H}_{(S^m, \mu^m)} \mid m \in \{1, \dots, N\}\}$. Let $S^U = \{s_m^U \mid m \in \{1, \dots, N\}\}$ and let $\mathcal{H}_{(S^U, \mu^U)}$ denote the common perceptual map. The *projection function* $\Pi : \mathcal{H}_{(S, \mu)} \cup \{\emptyset\} \rightarrow \mathcal{H}_{(S^U, \mu^U)} \cup \{\emptyset\}$ associates to a HLT $H = [s_i, s_j] \in \mathcal{H}_{(S, \mu)}$, the element $\Pi(H) = \pi(s_i) \sqcup \pi(s_j) = [s_{L_i}^U, s_{R_j}^U] \in \mathcal{H}_{(S^U, \mu^U)}$. For the empty set $\Pi(\emptyset) = \emptyset$.

This projection function extends the previous definition to non-basic HLTs.

3 A comparative study of news perception

In this section we study and compare news perception among four European countries, specifically, the UK, Germany, France and Spain, towards the Israel-Gaza war using the concepts presented in Section 2. The data set for this case is gathered from the GDELT data set [5].

In GDELT data set, the tone for an article is described in terms of six emotional dimensions: the average tone of the article, the positive score, negative score, percentage of words found in the tonal dictionary, percentage of active words, and percentage of self/group reference [9].

Although describing emotions would be more accurate by considering multidimensionality, this preliminary study specifically focuses on the negative score. The negative score represents the percentage of words conveying a negative emotional connotation. We have selected this dimension because negativity was the prevailing emotion in the tone of news during the specified period, from October 7th, 2023, to February 26th, 2024. This is largely attributed to the negative impact of the war on the news sentiment presented in European newspapers.

3.1 Methodology

The methodology of the use case follows seven steps:

1. *Define a baseline of linguistic terms for negative sentiment:* To do so we consider a previous period of time to define a baseline considering all the countries. We discretize the negative sentiment during this period and assign four levels of linguistic terms for the negative sentiment: "Low", "Medium", "High", and "Very high". Note that in this preliminary work, we have used a one-week period to establish the baseline. However, extending this period could be beneficial for capturing a broader range of sentiment nuances.
2. *Deduce countries' linguistic perceptual maps:* For each country, the linguistic perceptual map $\{(\mathcal{H}_{S^m}, \mu^m) \mid m \in 1, \dots, k\}$ is calculated based on the relative frequency with which newspapers within the country were associated with the four defined levels of negative sentiment. The motivation for assigning distinct linguistic perceptual maps to different countries lies in the varying usage of negative words across those countries.
3. *Obtain Common Perceptual Map:* From the perceptual maps determined for each country, $\mathcal{H}_{(S^m, \mu^m)}$, the common perceptual map, $\mathcal{H}_{(SU, \mu^U)}$ is calculated following Definition 3. The landmarks in the common perceptual map are renamed $\{\lambda_0, \lambda_1, \dots, \lambda_N\}$ for ease of reference and computation.
4. *Select articles associated with the concerned topic during the target period of time:* In this step, news referencing the concerned topic or actors involved are identified. Following GDELT terminology, an actor can be a person, country, geographical area, or organization closely related to the topic. News are filtered for either the topic or one of these types of actors.
5. *Represent their negative sentiment in its own linguistic perceptual map, and project it to the common perceptual map:* For each country, articles' negative sentiment during the defined period is represented in the specific linguistic perceptual map as "Low", "Medium", "High", or "Very high". Then we project articles' negative sentiment to the common perceptual map.

6. *Compute the centroid for each country per day:* For each day during the selected period, we compute, in the common scale, four different centroids corresponding to the central opinion of each country following Definition 5.
7. *Compare the news negativity among countries:* For each day we compute the distances from each country centroid to the term in the common scale that has the maximum level of negativity. Then we compare the results to analyze which country has the strongest level of negativity per day during the period.

3.2 Results

Articles were collected for a time period of 143 days, since October 7th, 2023, the day in which Israel attack started, to February 26th, 2024. GDELT Project translates into English those articles that were written in other languages.

Previously, as explained in the methodology, we define a baseline of linguistic terms considering all countries. The period considered was before the start of the war, specifically from 1st to 7th September, 2023. We discretized the negative sentiment during this period considering the most important newspapers of each country to determine the quartiles (see Table 1). In Table 2, the distribution of the four levels of linguistic terms for the negative sentiment: "Low", "Medium", "High", and "Very high" are presented.

Table 1. Set of thresholds obtained from 1st to 7th September, 2023

min	0
q_1	2.25
q_2	3.73
q_3	5.58
max	23.8

The perceptual map landmarks are calculated from the relative frequency of newspapers within the country associated with the four defined levels of negative sentiment. These relative frequencies determine the widths of the basic labels in the country's perceptual map.

Table 2. Distribution of the linguistic terms in the four countries.

Negative Tone	Germany	France	Spain	UK
Low	25%	20%	30%	22%
Medium	21%	30%	28%	24%
High	24%	26%	25%	26%
Very high	30%	24%	17%	28%

For each country, we use the relative frequency of the levels of negative sentiment (see Table 2) to define the landmarks in the partition associated with the linguistic perceptual map. The corresponding partitions of the unit interval, and their resulting perceptual maps, are the following for Germany (1), France (2), Spain (3), and the United Kingdom (4), respectively:

$$\begin{aligned}\mathcal{H}_{(S^1, \mu^1)} &: \{0.0, 0.250, 0.462, 0.698, 1.0\}; \\ \mathcal{H}_{(S^2, \mu^2)} &: \{0.0, 0.204, 0.496, 0.757, 1.0\}; \\ \mathcal{H}_{(S^3, \mu^3)} &: \{0.0, 0.303, 0.585, 0.829, 1.0\}; \\ \mathcal{H}_{(S^4, \mu^4)} &: \{0.0, 0.220, 0.457, 0.719, 1.0\}.\end{aligned}$$

Note that, for example, the language used in German newspapers tends to exhibit more extreme negative values than in the rest of the countries.

Next, the common perceptual map is obtained following Definition 4. The partition associated with the common perceptual map is:

$$\begin{aligned}\mathcal{H}_{(SU, \mu^U)} &: \\ \{0.0, 0.204, 0.220, 0.205, 0.303, 0.457, 0.462, 0.496, 0.585, \\ 0.698, 0, 719, 0.757, 0.829, 1.0\}.\end{aligned}$$

Note that the cardinals of S^1 , S^2 , S^3 and S^4 are equal to 4 in all countries, while the cardinal of S^U is $N = 13$ in this case.

Then, we calculate a centroid per each day and country within the common perceptual map. Finally, we compute their distances to the maximum value of the common perceptual map, i.e. s_{13}^U to numerically compare the negative sentiment of news among countries.

When comparing per each day distances to s_{13}^U among countries, the results show that there are significant differences among them. In the 56.5% of days during the period, Spain is the country with the most negative sentiment towards the war. In the 37% of days during the period, France is the country with the most negative sentiment towards the war. In the 5.8% of days during the period, United Kingdom is the country with the most negative sentiment towards the war. Only in one day during the period, Germany is the country with the most negative sentiment towards the war. It is noteworthy that this day, November 24th, 2023, marked the first day of the ceasefire in the war. As part of the agreement between Israel and Hamas, Hamas released 39 hostages, while Israel released 24 hostages on that day.

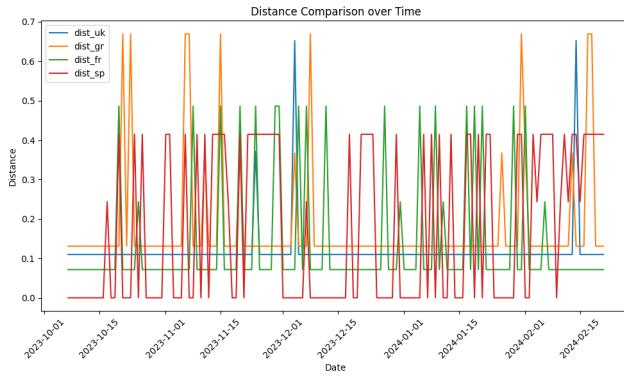


Figure 1. The distance to the maximum value of the common perceptual map per day cross all four countries during the first 143 days of the war.

Finally, for each country, we have computed the centroid corresponding to the complete period of time. Table 3 shows the centroids for each country computed in their respective linguistic perceptual maps, together with their expressions in the common perceptual map.

Table 3. Distribution of the sentiment centroids in their original perceptual maps and in the common perceptual map.

Centroids	Spain	France	UK	Germany
original maps	$\{s_4\}$	$\{s_4\}$	$\{s_4\}$	$\{s_4\}$
common map	$\{s_{13}^U\}$	$[s_{12}^U, s_{13}^U]$	$[s_{11}^U, s_{13}^U]$	$[s_{10}^U, s_{13}^U]$

Note that, considering the complete period of time, the top negative score s_4 is obtained for all the countries in their original linguistic perceptual map. However, in the common perceptual map, Spain is the only country among these four, that expresses the most significantly negative sentiment (the top negative score s_{13}^U).

4 Conclusions and future research

In this paper we present a methodology to aggregate sentiment coming from international newspapers. Unbalanced linguistic scales are considered to define different linguistic perceptual maps to characterize sentiment from news. We present a use case study focused on analyzing the negative sentiment of news coming from four European countries about the Israel-Gaza war. We conduct a comparative

analysis of news coverage across European countries with respect to the Israel-Gaza war, aiming to capture the negative sentiment towards this ongoing conflict.

Results show that there are differences in the negative sentiment among countries, where Spain is the country with the most negative sentiment not only considering the sentiment day by day but also all the period at once.

As a future work, we plan to study and compare the degree of consensus within each country as a measure of the existing polarization within news in each country. From the application point of view, we plan to consider data from the entire period of the Israel-Gaza war to study the dynamic evolution of sentiment towards the war. On the other hand, we plan to perform a multidimensional study taking into account the six emotional dimensions provided by GDELT.

Acknowledgements. This research has been partially supported by the PERCEPTIONS Research Project (PID2020-114247GB-I00), funded by the Spanish Ministry of Science and Information Technology.

References

- [1] W. Abuasaker, J. Nguyen, F. J. Ruiz, M. Sánchez, and N. Agell. Perceptual maps to aggregate assessments from different rating profiles: A hesitant fuzzy linguistic approach. *Applied Soft Computing*, 147: 110803, 2023.
- [2] O. Appel, F. Chiclana, J. Carter, and H. Fujita. Successes and challenges in developing a hybrid approach to sentiment analysis. *Applied Intelligence*, 48:1176–1188, 2018.
- [3] Z.-S. Chen, K.-S. Chin, L. Martinez, and K.-L. Tsui. Customizing semantics for individuals with attitudinal hflts possibility distributions. *IEEE Transactions on Fuzzy Systems*, 26(6):3452–3466, 2018.
- [4] K. D. Forbus. Qualitative modeling. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(4):374–391, 2011.
- [5] S. P. Leetaru K. Gdelt: global data on events, location, and tone, 1979–2012. *ISA Annual Convention*, 2:1–49, 2013.
- [6] C.-C. Li, Y. Dong, F. Herrera, E. Herrera-Viedma, and L. Martínez. Personalized individual semantics in computing with words for supporting linguistic group decision making: an application on consensus reaching. *Information Fusion*, 33:29–40, 2017. ISSN 1566-2535. doi: <https://doi.org/10.1016/j.inffus.2016.04.005>. URL <https://www.sciencedirect.com/science/article/pii/S1566253516300227>.
- [7] H. Liao, Z. Xu, and X.-J. Zeng. Distance and similarity measures for hesitant fuzzy linguistic term sets and their application in multi-criteria decision making. *Information Sciences*, 271:125–142, 2014.
- [8] J. Montserrat-Adell, N. Agell, M. Sánchez, F. Prats, and F. J. Ruiz. Modeling group assessments by means of hesitant fuzzy linguistic term sets. *Journal of Applied Logic*, 23:40–50, 2017.
- [9] J. Nguyen, A. Armisen, N. Agell, and Á. Saz-Carranza. Comparing global news sentiment using hesitant linguistic terms. *International Journal of Intelligent Systems*, 37(4):2868–2884, 2022.
- [10] O. Porro, N. Agell, M. Sánchez, and F. J. Ruiz. A multi-attribute group decision model based on unbalanced and multi-granular linguistic information: An application to assess entrepreneurial competencies in secondary schools. *Applied Soft Computing*, 111:107662, 2021.
- [11] L. Roselló, F. Prats, N. Agell, and M. Sánchez. Measuring consensus in group decisions by means of qualitative reasoning. *International Journal of Approximate Reasoning*, 51(4):441–452, 2010.
- [12] Y. Wu, Z. Zhang, G. Kou, H. Zhang, X. Chao, C.-C. Li, Y. Dong, and F. Herrera. Distributed linguistic representations in decision making: Taxonomy, key elements and applications, and challenges in data science and explainable artificial intelligence. *Information Fusion*, 65: 165–178, 2021.

Qualitative Modeling of Social Relationships

Kenneth D. Forbus

Qualitative Reasoning Group, Northwestern University
2233 Tech Drive, Evanston, IL, 60208 USA
forbus@northwestern.edu

Abstract

Social reasoning is a key capability in human cognition. Formalizing social reasoning can both improve our understanding of human cognition and support building AI systems that can perform it. The advantages of qualitative representations, such as abstraction of numerical values and compositional causal relationships between quantities, become especially important in domains where basic properties to formulate mathematical models are missing. However, social reasoning provides new challenges for qualitative reasoning, since, like many everyday reasoning problems, it involves fluently moving between discrete representations of events/actions and continuous causal models. This paper explores the hypothesis that the continuous aspects of social reasoning can be effectively modeled in qualitative process theory plus two extensions. These extensions are (1) incorporating discrete changes in the language of influences and (2) modeling aspects of episodic memory via sets of cases representing experience. We illustrate these ideas by formalizing aspects of social relationships involved in friendship.

Introduction

Social life is complicated. So much so that sociality is increasingly viewed as a driver of the evolution of intelligence (e.g. Tomasello, 2001). Thus understanding how social reasoning works can better help understand human cognition, as well as providing part of the foundation for creating AI systems that can understand our social world and even someday participate effectively in it.

We seem to think about some aspects of social life in continuous terms. For example, we can talk about events bringing us closer to someone, and gauge which of our friends is more likely to be relied on in a tough situation. However, as with many everyday phenomena, the level of precision in available information is a mismatch with the requirements of traditional mathematical modeling methods. To gather numerical data requires having some notion of units, for example – when we speak of two people being close, how would we quantify that as a number? There have been attempts to model social relationships mathematically, but there is little quantitative data upon which to base such models, nor constraints on their internal parameters. Thus formalisms developed in the qualitative reasoning community offer a way to build models that are closer to what kinds of information are actually available about the phenomena.

However, social reasoning raises interesting challenges for qualitative reasoning. It requires drawing conclusions from experiences in the everyday world, which means incorporating rich representations of events, their participants, and relationships among them. It requires shifting between continuous and discrete models of actions, as when one takes a series of discrete interactions in the aggregate to approximate derivatives over intervals. And it requires estimating properties of interactions over experiences, hence addressing properties of episodic memory.

The rest of this paper describes an initial attempt to model the continuous aspects of social relationships qualitatively, using qualitative process theory (Forbus, 1984;2019) with two extensions. We begin by providing some background both about social psychology and some ontological assumptions we build upon for handling events, finite symbol quantity values, sets, and cases. Then we describe the two extensions to QP theory. The first is to re-purpose a discrete notion of influence due to Kim (1993) to handle causal reasoning about the occurrence of events on continuous parameters. The second is a very simple formalization of episodic memory and how to connect changes in events to changes in qualitative values. Next, we describe a basic encoding of social relationships, focusing on continuous properties. An extended example shows how these ideas can be used together to do at least one aspect of social reasoning. We close with conclusions and future work.

Background

We begin with the aspects of social psychology we are drawing upon, then the ontological assumptions in addition to QP theory that we need to make.

Some Social Psychology

Given the complexity of human social life, it is perhaps unsurprising that social psychology has not settled on a single, universally agreed upon theory of social relationships, let alone a formal version of such a theory. There are multiple frameworks with varying degrees of adoption. For example, Fisk (1992, 2004) argues that social relationships can be broken down into four categories. The first, communal sharing, focuses on what people have in common. This includes being members of the same family, workplace, club, and so on. The second, authority ranking, describes interactions in terms of ordered differences, such as seniority, age, gender,

or caste. The third, equality matching, uses balances of contributions, such as turn-taking and exchanges of favors. The fourth, market pricing, uses money or some implicit continuous parameter to evaluate interactions. This includes rents, dividends, interest rates, and evaluating relative benefits of a relationship to those involved. Each of these categories is described as modes of interactions, “mods” in Fisk (2004), which can be combined with culturally specific prototypes and patterns, “preo” in Fisk (2004), to describe the practices of a group. In compositional modeling terms, mods and preos are analogous to model fragments, with preos modifying mods, such that a situation model composed of such model fragment instances would be a model of how that cultural group operates. While examining this analogy more closely could be productive, we focus here on the effects of interactions on individuals and their relationships, rather than constraining what interactions people will have, which is more in the territory of social norms (Malle et al. 2019; Olson & Forbus, 2021).

While Fisk’s account focuses on providing a mechanism, other social psychology models such as Kelley et al. (2003) focus more on cataloging the phenomena. Kelley et al. argue that the construct of situation is central in social psychology, because the situations that social beings find themselves in are major factors in determining what they do. Thus classifying types of situations serves the purpose of carving the phenomena up into units amenable for analysis. Kelley et al. (2003) argues that recognizing such situations is an important force in our evolution. Formalizing these situations, again, would be an interesting enterprise, but would take us far beyond what a QR focus can provide.

To provide an initial focus for modeling, we build on the account of friendship due to Rawlins (1992). Rawlins’ account is informal and descriptive, not mechanistic. He argues that friendship is a kind of social relationship that is (1) voluntary, (2) negotiated by both parties involved, (3) provide some sense of equality for the parties involved, (4) require mutual involvement, and (5) have an affective component¹. The last three characteristics are promising for qualitative modeling because they seem to involve continuous factors. For example, one factor in equality is that the needs and desires of both friends are important, making large or long-duration imbalances something important to detect. Similarly, mutual involvement is measured by both parties being willing to spend time together at an appropriate amount and frequency. Finally, pairs of friends are closer to each other than people who are not friends. They know each other’s histories and build up considerable shared history. The notion of closeness is analyzed below, since it seems central to friendship and social relations.

Rawlins also proposed a seven-stage model of the trajectory of friendship, encompassing both its growth and decline. The first stage are *role-delimited interactions*, e.g. the relationships you have with other people when you are shopping, driving, mentoring, etc. Within such roles, the next step involves what Rawlins calls *friendly relations*, where there is more mutual disclosure beyond what the roles require. Then comes the initiation of interactions outside the roles, a stage Rawlins calls *moves-toward-friendship*. Then comes *nascent friendship*, where interactions are no longer following role stereotypes at all, and norms for what and

how to communicate are established for the relationship. Some topics might be declared out of bounds, (e.g. religion or politics), and continued compliance with those mutually agreed-upon norms is part of the process of deepening trust. Interactions with others start to take the friendship into account as well. At some point nascent friendships become *stabilized friendships*, where interaction patterns and norms are stable and mutually agreeable. Rawlins observed that stabilized friendships fall into three types: *active friendships* involve regular mutual interactions currently, versus *dormant friendships* where mutual interactions have tapered off, although it could be quickly restarted by interacting again. Finally, *commemorative friendships* are those where the bulk of the interactions were in the past, with only minimal current interactions. The last two stages that Rawlins identified are involved in the dissolution of friendship. A *waning friendship* is one that starts to decline in its importance in our lives. The sources of waning might be a reduction in closeness, due to one or both parties not investing enough in events that sustain friendship, or negative events like betrayals. Finally, in the post-friendship phase, the friendship continues to provide memories that influence future relationships, e.g. great activities to have done, experiences best avoided.

Being informal, there are many open questions in Rawlins’ account. For example, how can one tell a post-friendship from a dormant stabilized friendship? While model fragments might be used to encode some aspects of these various stages, the criteria that should be used to define limit points to transition between them is far from clear. Nevertheless, building up formal qualitative models of such theories might both provide better social reasoning for AI systems, as well as perhaps helping produce more formal but still qualitative social psychology theories. Such formal qualitative models would enable the generation of testable predictions, while at the same time being better suited to the kinds of evidence and data available, compared to traditional mathematical models. This is not unprecedented, as the work of Bredeweg et al. (2008) in ecology and de Jong (2008) in genetic regulatory networks illustrate. Here we will start small and build up a simple model of social relationships that expresses some continuous aspects of friendship.

Ontological Assumptions

For the aspects of this model that lie outside QP theory, we freely draw upon the OpenCyc ontology, as used in NextKB². OpenCyc is a subset of Cycorp’s Cyc KB (Lenat et al. 1990) contents that is freely available³. It provides a broad commonsense ontology which can be linked to QP theory constructs and is grounded in natural language (Forbus, 2023), making it useful for describing the open-ended nature of events and relationships in the world. For example, there are 16,715 subcategories of Event, and 2,643 distinct role relationships that express their properties. This broad vocabulary is useful given the nature of human social life.

The NextKB ontology provides support for Hayes’ (1985) notion of histories, where change is represented in terms of bounded pieces of space/time whose properties

¹ This summary draws upon Wrench et al. (2023)’s account of Rawlins’ work.

² <https://www.qrg.northwestern.edu/nextkb/index.html>

³ Creative Commons Attribution license.

vary along those axes. In OpenCyc, the kinds of things that can have histories are instances of `SpatialThing-Localized`, which inherits from `TemporalThing` and `SpatialThing`. The predicate `holdsIn` provides a modal operator that specifies that, over a given temporal extent denoted by a `TemporalThing`, a given proposition holds. This provides a means of specifying what properties are true during an event, for example. The logical function `AtFn` is used to denote the spatiotemporal slice of an entity during a subset of a history, e.g. the speed of a falling object is higher at a point later in its trajectory than at its start.

The Cyc ontology supports microtheories (Guha 1992), a form of context. All reasoning is performed with respect to some microtheory and those microtheories it inherits from (via the `genlMt` relation). Qualitative states and models of the contents of the minds of others are both implemented via microtheories, for instance.

NextKB supports the traditional QR notion of a quantity as a fluent, which takes different values at different times. In addition to ordinal and signs, NextKB inherits a well worked out ontology of symbolic values and numerical properties from OpenCyc. Properties such as `Happiness` are ontologized so that instances of them are values, either symbolic or numerical. The symbolic values include the kind of finite symbol algebra commonly found in QR. Logical functions (e.g. `HighAmountFn`) are used to provide a general approach to specifying such values. OpenCyc also provides a rich collection of units and conversions between them. Units are represented by logical functions, so that non-atomic terms like `(HoursDuration 3)` bundle units with values.

The preferences of an agent can be expressed via `prefers`, a ternary predicate taking an agent and two sentences, meaning that the agent prefers situations in which the first sentence is true over those in which the second sentence is true.

NextKB supports both intensional and extensional representations of sets. An operator is provided for gathering the bindings satisfied by a conjunction of statements (`the-ClosedRetrievalSetOf`) evaluated over an extensionally specified set. We will use this operation in modeling operations over episodic memory below.

Extensions to QP Theory for Social Reasoning

QR has mostly focused on continuous processes, but there have been interesting exceptions. Simmons (1983) described a notion of discrete process to handle reasoning about geological processes, which happen too slowly to be directly observed (earthquakes and volcanos excepted), but whose occurrences over historical time are important to understand. His representation of change used operations on a diagrammatic representation of layers under the Earth. The representations below are typically also considered discrete events, and while a more physically grounded event representation would include a spatial/diagrammatic component, we do not do this here. Integration of QP theory with discrete planning has been done via STRIPS operators in environments (Forbus, 1989), compiling processes into operators for a temporal planner (Hogge, 1987), and more tightly integrated with plan operators (Drabble, 1993). The extension here to handle discrete effects in events is closest to Hogge (1987), but the representations are used differently in

reasoning. The second extension concerns modeling episodic memory. Some prior work has explored the use of QR in the representation of episodic memories (Hancock & Forbus, 2021), but not in constructing a formal model of episodic memory per se.

Handling Discrete Effects in Events

Consider a pleasant outing undertaken by two friends, Pat and Kit. They walked through the woods, picnicked in a clearing, and swam in a lake. Each of these constituent events of the outing could be decomposed further into sub-events as needed. Some of those sub-events can, in turn, be viewed as including occurrences of continuous processes, such as walking and swimming. In other cases, the subevents may best be viewed as discrete events. For example, the picnic consists of particular events of eating, drinking, and conversing, bookended by setting up the picnic and cleaning up afterwards. But some of these events can be construed in terms of continuous processes. Thus the event of drinking a glass of wine can be decomposed if needed into a set of movements of the glass/liquid combination, the pouring of the liquid, and so on. To infer the causal import of events, we need to combine continuous and discrete models of effects across multiple levels of events. For example, given the wine consumption at the picnic, are either Pat or Kit in shape to go swimming? Answering this question does not require fine-grained decomposition of constituent events, only knowing that wine was consumed, and some means of estimating how much. This is an example of a key problem in commonsense reasoning: determining how to compute effects of events and processes across levels of description without getting bogged down in irrelevant details.

For continuous processes, we use the qualitative mathematics of influences from QP theory. That is, qualitative proportionalities (`qprop+`, `qprop-`) provide representations of partial information about algebraic causal connections (e.g. `(qprop+ (level (ContainedLiquid wine glass)) (mass (ContainedLiquid wine glass)))`). Direct influences (`i+`, `i-`) represent partial information about integral causality (e.g. `(i- (energy Kit) (rateFn (Walking Kit)))`). For discrete changes, we adopt H. Kim's (1993) extensions originally developed to encode abrupt changes:

`(increase <qty>)` indicates that `<qty>` increases over the interval of interest

`(decrease <qty>)` indicates that `<qty>` decreases over the interval of interest

`(increaseBy <qty> <amt>)` indicates that `<qty>` increases by `<amt>` over the interval of interest

`(decreaseBy <qty> <amt>)` indicates that `<qty>` decreases by `<amt>` over the interval of interest.

Unlike `i+,i-`, these relationships make no specification as to the derivative of `<qty>` at any particular sub-interval for the interval of instance, they only concern the net effect across the interval in question. In Kim (1993) this was used, for instance, to model the effects of combustion in a four-cycle engine, which is for some purposes modeled as an impulse. The same ambiguity regarding when within an interval that a change happens is used here for intervals covering substantial intervals of time, e.g. a picnic.

Returning to our picnic example, `(increaseBy (Wine-ConsumedFn Kit) (GlassesFn 3))` states that, in the event for which this statement appears, Kit consumed three glasses of wine. If `P1` denotes the picnic, then

```

(holdsIn P1
  (increaseBy (WineConsumedFn Kit)
    (GlassesFn 3)))

```

Such a conclusion might be reached by combining the amount of wine Kit drank across the entire picnic, i.e.

```

(evaluate ?n-drinks
  (TheClosedRetrievalSetOf ?drinks
    (and (occursDuring ?sub-e P1)
      (isa ?sub-e DrinkingEvent)
      (doneBy ?sub-e Kit)
      (substanceConsumed ?sub-e Wine)
      (amountConsumed ?sub-e (GlassesFn ?n))
      (unifies ?drinks (GlassesFn ?n))))))

```

The extraction of the total number of glasses consumed is straightforward.

So far, we have focused on the physical aspects of Kit and Pat's outing. But what are the social effects? Again, we only have finite qualitative symbol systems and ordinal relationships to express preference information. This means results will often be ambiguous, but that is the nature of qualitative reasoning. A social reasoner must evaluate as best it can the effects on the participants in terms of whether or not it is a positive, negative, or mixed experience. This requires taking their preferences into account. We do not assume perfect information. Each social reasoner must do the best that it can with the information it has.

There are four representations of preferences to consider between two people: Their own beliefs about what each of them prefers, and their own beliefs about what the other prefers⁴. These can be kept distinct via microtheories, e.g. (*PrefersBeliefsOfFn* Kit Pat) is the set of preferences that Pat believes that Kit has. There are multiple dimensions that could be compared for evaluative purposes, e.g. fiscal requirements, physical stamina, etc. and more likely this set is simply a subset of the larger set of beliefs that Pat has about Kit. We stick with preferences here for simplicity.

Reasoning about Change in Sets of Events

Our intuitions about friendship, as well as Rawlin's model, tell us that our feelings are determined in part by our recollections of shared experiences with our friend. Thus we need to have a way to model the relevant types of memories and how they change over time. We are focused here on episodic memories, specifically memories of events that an agent has participated in with other people, making them grist for building/evaluating that social agent's relationship with that other (or others). We will represent each memory via a case (implemented as a microtheory) whose contents are one or more occurrences of events. For example, two friends having a dinner out can be described in terms of such a microtheory.

People's memories are personal, subjective and noisy. The same event (or network of interlocking events, without loss of generality) might be remembered in very different ways by the participants⁵. We use the relationship (*episodicMemoryOf* <mt> <Person>) to indicate that the microtheory <mt> of events is part of <Person>'s episodic memories.

⁴ In the early stages of a relationship, preferences can be misstated in order to increase closeness, e.g. expressing interest in board games early on but then refusing to play once married <https://www.nytimes.com/2024/05/31/magazine/judge-john-hodgman-on-compulsory-game-nights.html>

For each Person, there is a microtheory denoted by (*MemoriesOfFn* <Person>) whose contents are that person's memories. Thus *episodicMemoryOf* statements in that person's *MemoriesOfFn* microtheory indicate that that person does indeed have that memory. *episodicMemoryOf* statements in someone else's *MemoriesOfFn* microtheory represent someone else's belief as to what that person remembers about an experience. We assume that there are memories beyond episodic memories in *MemoriesOfFn*. For our purposes, we include other narratives (e.g. stories that a person has understood from conversation, reading, and watching), but not semantic memory or skill memory, because these are not used in similar ways in social reasoning. The episodic memories of an agent (i.e. the extension of *episodicMemoryOf* statements) will be denoted using *EpisodicMemoryFn*.

As people gain experience, and as their knowledge about their experiences changes (learning more about a party, forgetting a slight that you perceived), this can lead to changes in social relationships. Such changes are not tied to the underlying parameter changes as directly as the normal qualitative mathematics of influences, nor are they continuous integrations anymore. The notion of derivative across intervals of time representing occurrences of discrete events is still useful even though it is more granular. We define the usual signs of derivatives in terms of ordinal relationships over a quantity being tracked that is affected by an event. This means we need to extend ordinals across sets of events, and extend changes to include both changes in parameters (being updated about an event, or forgetting aspects of an event).

We define the extension of a quantity Q over a set of episodic memories M as (*quantityAspect* Q M). Thus we can say

```

(> (quantityAspect Enthusiasm
  (subsetOfType Snorkeling
    (EpisodicMemoriesOf K)))
  (quantityAspect Enthusiasm
    (subsetOfType DentalWork
      (EpisodicMemoriesOf K))))

```

Evaluating such ordinal relationships is easy when there are numerical values, as can sometimes be done with physical domains. That is not an option here, but fortunately QR research has provided some useful ways of eking out conclusions from partial information. For example, bipartite graph partitioning of opposing signs used in influence resolution can sometimes generate answers when enough ordinal relationships between specific events are known. Similarly, symbolic algebras can often provide ordinal information across a broader range of quantities. For instance, if (*VeryHighAmountFn* Enthusiasm) applied to Snorkeling and (*VeryLowAmountFn* Enthusiasm) applied to DentalWork, the ordinal above would follow from these.

A particularly subtle effect is when a value has increased, but not so much as to be detectable via ordinals computed over aspects. Enough such small increases can lead to a derivative increase if the comparison were across

⁵ See Kurosawa's Rashomon for an extreme example. <https://en.wikipedia.org/wiki/Rashomon>

a broader range of memories, as when values “break over” in order of magnitude representations (Dauge, 1993). A similar mechanism might be useful here.

A Simple QP Model of Social Relationships

Now we have enough representational machinery built up to construct a simple model of social relationships. We start by formalizing the concept of a social relationship between two people. We formalize social relationships as conceptual entities, represented by model fragments. Each involves a pair of instances of `Person`⁶.

```
(defModelFragment SocialReln
  :participants ((?me :type Person)
                 (?other :type Person))
  :conditions ((knowEachOther ?me ?other))
  :consequences ((hasQuantity
                  (ClosenessFn ?self))
                 <...>))
```

The variable `?self` is a meta-linguistic convention: `?self` is always bound to the model fragment instance, so that its properties can be specified in the definition of the model fragment type. More consequences of this model fragment are enumerated below.

Recall that in model fragments, the variables for participants define role relations. So if Pat knows Kit, then

```
(isa SR1 SocialReln)
(me SR1 Pat)
(other SR1 Kit)
```

Notice that this relationship is unidirectional – if `knowEachOther` is a symmetric predicate, then the first implies a social relationship in the other direction:

```
(isa SR2 SocialReln)
(me SR2 Kit)
(other SR2 Pat)
```

This allows for asymmetric relationships, e.g. Pat might feel more close to Kit than the other way around. By building in a perspective `SocialReln` also better supports reasoning from that perspective, e.g. an agent (`?me`) reasoning about whether someone else (`?other`) might be called upon to help or participate in some other mutual activity.

We will take the closeness that a person feels for another (i.e. `(ClosenessFn SR1)` as how close Pat feels to Kit) as a quantity that, when sufficiently high, causes them to believe that the other person is a friend. In other words, there is a limit point at which this transition happens, but we will not specify it except to denote it as `(FriendLimitPointFn SR1)`.

Closeness appears to depend on multiple factors. One aspect is shared background, e.g. if two people are routinely engaged in the same kinds of activities and have overlapping social networks, they have a built-in basis of common ground. Another aspect is shared experiences, which builds

up a shared history together. Notice that this shared history does not always have to be enjoyable: A slogan in the US military is “Shared pain leads to unit cohesion.” This assumes the source of the shared pain is outside the dyad, e.g. an unexpected thunderstorm ruining a picnic might bring people closer together, especially if they worked together to ameliorate its negative effects. On the other hand, if one of the planners was feckless and ignored a weather forecast predicting a serious thunderstorm, that would most likely decrease, not increase, closeness.

To capture shared experiences, we model episodic memory as a set of cases. Recall that

`(EpisodicMemoryFn <agent>)`

denotes the set of cases that constitute the episodic memory of agent `<agent>`. We define a subset of episodic memories relevant to a social relationship by those which mention another Person as

`(InteractionEpisodicMemoriesFn <agent> <other>)` and those episodic memories relevant to a social relationship via

`(SocialRelnMemoriesFn <SocialReln>)`

consisting of the `InteractionEpisodicMemoriesFn` for the me and other of that relationship.

We split positive and negative aspects of events because people seem to track them separately. For example, we can distinguish between an event whose net impact is small because it only had a small positive impact, or because there are large impacts of opposite signs, the latter being a more fraught situation. The positive and negative effects of events involving the people in a social relationship will be represented by two quantities, `(PosExperienceFn ?self)` and `(NegExperienceFn ?self)`. These are accumulations over the set of episodic memories of the me of the relationship for memories where the other is involved. Thus when a new event is experienced, its positive and negative impacts will be considered in estimating these quantities. Recall that there are no numerical values associated with these quantities, by assumption. Instead, ordinal values are updated based on local information. Suppose a new event `E'` is added to episodic memory, and it was more positive than negative. Then, treating closeness as an extensive parameter, the new value for closeness, whatever it is, is higher than the value before this event. The two people have become closer. Should it have been more negative than positive, the new value for closeness would be recorded as less than the prior value, i.e. $D_s = -1$.

This method of tracking causal changes due to differences caused by adding events⁷ relies on local changes and the accumulation of ordinals across time. This detailed record-keeping may or may not be psychologically plausible, and it does not provide easy comparison across people, e.g. are you closer to one friend than another? There might be a summarization mechanism that tracks accumulation of ordinal changes through changes in a parallel symbolic algebra representation, e.g. `HighAmountFn` transitioning to `VeryHighAmountFn`. We discuss possible quantitative extensions in future work.

⁶ In OpenCyc, `Person` inherits from `IntelligentAgent` and `SocialBeing`, but does not include organizations, hence apt for this purpose

⁷ Exactly how much forgetting (as opposed to failing to consciously retrieve) occurs in episodic memory is still an open question. If episodes are

forgotten, does that process somehow update the quantities that it was involved in changing? This seems unlikely.

We introduce the following additional quantities for `SocialRein`, with the understanding that this set is likely incomplete:

- `TrustLevelFn`: How much you trust the other person with regard to information-sharing.
- `ReliabilityFn`: How likely will they do what they say they will do.
- `HelpfulnessFn`: How likely they are to be willing to help do something.
- `InterestsOverlapFn`: How many of your interests do they share?
- `FriendsOverlapFn`: How much do your social networks overlap?

What do these parameters depend upon? Let us start with trust. The common business metaphor “trust is built in drops and lost in buckets” suggests an accumulation, albeit asymmetric in flow rates, which would make it an extensive parameter. For each event `E` in `InteractionEpisodicMemoriesFn` for a social relationship `S`, for every information sharing norm `N` in `S`, either `E` is agnostic with respect to `N`, or represents compliance with `N`, or represents a violation of `N`. Depending on the relationship between `E` and `N`, `E` either doesn’t contribute to either experience parameters, contributes to `(PosExperienceFn S)`, or contributes to `(NegExperienceFn S)`. In other words, adhering versus violating norms on information sharing in the relationship should increase/decrease closeness as well as trust. A separate `PosTrustRateFn` and `NegTrustRateFn` are introduced to represent the effects of adherence/violation to trust, with the magnitude of `NegTrustRateFn` being much larger than `PosTrustRateFn`, to model the gradual accumulation of trust and sharp dissolution of it upon betrayal. (An order of magnitude relationship might not be amiss here.) Note that, in addition to asymmetry, keeping these rates separate from their effect on closeness should better enable modeling that someone might get closer again to another while no longer trusting them with regard to keeping secrets.

Reliability can be modeled as a ratio of commitments honored to commitments made, e.g. the consequences of `SocialRein` should include

```
(qprop+ (ReliabilityFn ?self)
        (NCommitmentsHonoredFn ?self)
(qprop- (ReliabilityFn ?self)
        (NCommitmentsMadeFn ?self))
```

Where `NCommitmentsMadeFn` and `NCommitmentsHonoredFn` are the cardinalities of sets whose members consist of the set of commitments made by the other and the number of those which were honored, respectively. This illustrates the importance of reputation: An agent must assess this information either directly, from experiences with the other, or indirectly, from what yet other agents say about the other agent in the relationship. That said, there are commitments with varying levels of importance, e.g. keeping secret a surprise birthday party versus keeping secret that someone plans to leave their job. So cardinality is unlikely to be sufficient to capture the causal relationship here. This is not an isolated case, so we should develop a general representation that can be specialized via reasoning appropriately. Let us define `ImportanceScoreFn` as a binary function whose first argument is a continuous quantity and whose second argument is a social agent. Its range is in turn a unary function

whose domain is a set of events and whose range is a set of quantity values. For example,

```
((ImportanceScoreFn eventValue <me>
    (EventsWithCommitments <other> <me>)))
```

provides a set of quantity values for the set of events,

where `(EventsWithCommitments <other> <me>)` expands to

```
(TheClosedRetrievalSetOf ?e
  (and (isa ?e Event)
       (commitmentInEvent ?e <other>)
       (eventValue ?e ?v)))
```

and `(EventsCommitmentsHonored <other> <me>)` is

```
(TheClosedRetrievalSetOf ?e
  (and (isa ?e Event)
       (commitmentInEvent ?e <other>)
       (commitmentHonoredInEvent ?e <other>)
       (eventValue ?e ?v)))
```

We are assuming finite symbol values here for simplicity, accumulating ordinal information about particular properties of events is also doable but somewhat more complex.

Helpfulness can be defined analogously, i.e. a positive influence based on the number of times the other helped the me agent by some action, including joint activity, plus a negative influence based on the perceived number of opportunities to be helpful that the other agent had but did not take. Again this could be based simply on cardinality, the number of times they were helpful, or each contribution could be scaled based on utility, e.g. giving someone \$5 versus giving someone your kidney for a transplant.

The final two quantities that seem relevant are the overlaps in interests and in friends. Again, these might simply be cardinality in set intersections or scaled based on significance. For example, in the US it was once not uncommon for spouses to belong to different political parties, whereas now political orientation is typically a gating factor on long-term involvement. Similarly, if one shares close friends with someone else, that is likely to have more impact on closeness than sharing assorted random acquaintances.

This formulation of sets and importance measures to define quantity values is quite different from the traditional notion of directly influenced parameters, where the derivative of a quantity is specified continuously over time. In traditional continuous change models, effects accumulate continuously. For these parameters, the values change discretely, as the members of particular sets of events, interests, or people change over time (and perhaps change in evaluation as well). Nonetheless, the compositional causal relationships do seem to capture the intended effects of changes in the sets and in the evaluations of members of those sets.

Example: Planning an Outing

Pleasant outings often involve planning, which should take into account the preferences of the people involved. Suppose Kit did the planning for the outing described above and has already decided to propose a picnic and a swim. There

are two paths to the clearing, one a pleasant amble and another requiring climbing equipment. Suppose further that

```
(PrefersBeliefsOfFn Kit Kit):
(attitudeTowardsType Kit Walking
    (HighAmountFn Enthusiasm))
(attitudeTowardsType Kit RockClimbing
    (VeryHighAmountFn Dislike))
(PrefersBeliefsOfFn Kit Pat):
(attitudeTowardsType Pat Walking
    (MediumAmountFn Enthusiasm))
(attitudeTowardsType Pat RockClimbing
    (HighAmountFn Enthusiasm))
(attitudeTowardsType Pat Swimming
    (HighAmountFn Enthusiasm))
```

We assume that these `attitudeTowardsType` statements are generated via computations over episodic memories, both of things that they have done together but also Kit's understanding of Pat's self-reports or third-party stories about Pat.

Given straightforward reasoning about relative magnitudes and the negative relationship between enthusiasm and dislike, we get a conflict in preferences:

```
(prefers Pat (activityInPlan P1 RockClimbing)
    (activityInPlan P1 Walking))
(prefers Kit (activityInPlan P1 Walking)
    (activityInPlan P1 RockClimbing))
```

where `activityInPlan` means that the plan denoted by the first argument includes one or more instances of the concept denoted by the second argument. So while Kit believes that Pat would prefer rock climbing, Kit does not want to be miserable, which would make the outing less fun for them both, and so proposes walking, which both should find acceptable.

Differences in beliefs can lead to surprises:

```
(PrefersBeliefsOfFn Pat Pat):
(attitudeTowardsType Pat Walking
    (MediumAmountFn Enthusiasm))
(attitudeTowardsType Pat RockClimbing
    (HighAmountFn Enthusiasm))
(attitudeTowardsType Pat Snorkeling
    (HighAmountFn Enthusiasm))
(attitudeTowardsType Pat Swimming
    (LowAmountFn Enthusiasm))
```

Kit may have inferred their belief that Pat is very enthusiastic about swimming from hearing that Pat is a snorkeling enthusiast. But if Pat's love of snorkeling comes from seeing coral reefs, swimming in a lake just isn't the same thing, hence the low enthusiasm for swimming per se. Nonetheless, it would still be a net positive experience for Pat, and hence his acceptance of the proposal.

Conclusions and Future Work

Like many other aspects of commonsense, social reasoning seems to have continuous aspects, and this paper argues that QP theory with two extensions may be able to formalize those aspects. These extensions bridge from qualitative

modeling to the more discrete world of events and the accumulation of these events into episodic memories that are analyzed to track relationship parameters over time. This account relies on higher-order representations, e.g. microtheories for representing cases and states of belief and states of affairs in the world.

There are two next steps. The first is to expand the formalization to handle more phenomena. For instance, if someone is betrayed, how does that impact their interpretation of their episodic memories of prior interactions? Are the continuous aspects of people's models stored in a distributed fashion (e.g. Friedman et al. 2018), so that different models for another person are retrieved under different classes of situations? The second is to implement the non-QP aspects of the reasoning described here, to test these ideas at reasonable scale. We plan to explore whether or not this account can be extended to support story understanding, e.g. to predict changes in social relationships between characters as the events of a story unfold.

There is a looming open question: How far can a purely qualitative account go, especially as the size of episodic memory grows? Is a quantitative substrate inevitable, to facilitate cross-person comparisons? Models of emotion, for example, compute appraisal variables (Gratch & Marsella, 2004; Wilson et al. 2013), and similar computations could be used for the quantities used here. It could be that some internal quantities are used to track the impact of experiences, but that qualitative representations are used to facilitate planning and prediction. This is a question worth exploring.

Acknowledgments

This research was sponsored by the US Air Force Office of Scientific Research under award number FA95550-20-1-0091.

References

- Bredeweg, B., Salles, P., Bouwer, A., Liem, J., Nuttle, T., ... Zitek A. (2008). Towards a structured approach to building qualitative reasoning models and simulations. *Ecological Informatics*, 3,1-12.
- Dauge, P. (1993). Symbolic reasoning with relative orders of magnitude. *Proceedings of the Thirteenth IJCAI*, pp. 1509-1515. Chambery, France.
- de Jong, H. (2008). Qualitative modeling and simulation of bacterial regulatory networks. In A. Uhrmacher & M. Heiner (Eds.), *Computational Methods in Systems Biology*. Springer-Verlag, Berlin, Germany.
- Drabble, B. (1993). EXCALIBUR: A program for planning and reasoning with processes. *Artificial Intelligence*, 6291:1-40.
- Fisk, A. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review*, 99, 689-723.
- Fisk, A. (2004). Relational Models Theory 2.0. In Nick Haslam (Ed.) *Relational Models Theory: A Contemporary Overview*. LEA Associates, New Jersey.
- Forbus, K. (1984). Qualitative process theory. *Artificial Intelligence*, 24, 85-168.

- Forbus, K. (1989). Introducing actions into qualitative simulation. *Proceedings of the 11th IJCAI*, pp. 1273-1279, Detroit, MI.
- Forbus, K. (2019). *Qualitative Representations: How People Reason and Learn about the Continuous World*. MIT Press, Cambridge, MA.
- Forbus, K. (2023) Domain Theories for Commonsense Reasoning from Language-Grounded Ontologies. *Proceedings of the 36th International Workshop on Qualitative Reasoning* Krakow, Poland.
- Friedman, S., Forbus, K., & Sherin, B. (2018). Representing, Running, and Revising Mental Models: A Computational Model. *Cognitive Science*, 1110-1145. DOI:10.1111/cogs.12574.
- Gratch, J. & Marsella, S. (2004). A domain-independent framework for modeling emotion. *Journal of Cognitive Systems Research*, 5(4):269-306.
- Guha, R. (1992) *Contexts: a formalization and some applications*. PhD. Thesis, Stanford University.
- Hancock, W. & Forbus, K. (2021). Qualitative Spatiotemporal Representations of Episodic Memory for Strategic Reasoning. *Proceedings of QR-2021*.
- Hayes, P. (1985). Naïve Physics I: ontology for liquids. In J.R. Hobbs & R. Moore (Eds.), *Formal Theories of the Commonsense World*. Ablex: Norwood, NJ.
- Hogge, J. (1987). Compiling plan operators from domains expressed in qualitative process theory. *Proceedings of AAAI 1987*, AAAI Press, Menlo Park, CA.
- Kelley, H., Holmes, J., Kerr, N., Reis, H., Rusbult, C., & Van Lange, P. (2003). *An Atlas of Interpersonal Situations*. Cambridge University Press, Cambridge, England.
- Kim, H. (1993). *Qualitative reasoning about fluids and mechanics*. Ph.D. dissertation and ILS Technical Report, Northwestern University.
- Lenat, Douglas B.; Guha, R. V.; Pittman, Karen; Pratt, Dexter; Shepherd, Mary (August 1990). "Cyc: Toward Programs with Common Sense". *Communications ACM*. 33 (8): 30–49. doi:10.1145/79173.79176
- Malle, B.F.; Bello, P.; and Scheutz, M. (2019). Requirements for an Artificial Agent with Norm Competence. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 21-27).
- Olson, T. & Forbus, K. (2021). Learning Norms via Natural Language Teachings. *Proceedings of the 9th Annual Conference on Advances in Cognitive Systems 2021*.
- Rawlins, W. (1992). *Friendship matters: Communication, dialectics, and the life course*. Routledge, 2017.
- Simmons, R. (1983). *Representing and reasoning about change in geologic interpretation*. Ph.D. dissertation, MIT, Cambridge, MA.
- Tomasello, M. (2001). *The Cultural Origins of Human Cognition*. Harvard University Press, Cambridge, MA.
- Wilson, J., Forbus, K., and McClure, M. (2013). Am I Really Scared? A Multi-phase Computational Model of Emotions. *Proceedings of the 2nd Conference on Advances in Cognitive Systems*.
- Wrench, J., Punyanunt-Carter, N., & Thweatt, K. (2023) Chapter 10: Friendship Relationships, in *Interpersonal Communication: A Mindful Approach to Relationships*, LibreTexts. <https://milnepublishing.geneseo.edu/interpersonalcommunication/chapter/10/>

Unveiling Ontological Commitment in Multi-Modal Foundation Models

Mert Keser^{a,b,*1}, Gesina Schwalbe^{c,**1}, Niki Amini-Naieni^d, Matthias Rottmann^e and Alois Knoll^b

^aTechnical University of Munich, Germany

^bContinental AG, Germany

^cUniversity of Lübeck, Germany

^dUniversity of Oxford, UK

^eUniversity of Wuppertal, Germany

Abstract. Ontological commitment, i.e., used concepts, relations, and assumptions, are a corner stone of qualitative reasoning (QR) models. The state-of-the-art for processing raw inputs, though, are deep neural networks (DNNs), nowadays often based off from multimodal foundation models. These automatically learn rich representations of concepts and respective reasoning. Unfortunately, the learned qualitative knowledge is opaque, preventing easy inspection, validation, or adaptation against available QR models. So far, it is possible to associate pre-defined concepts with latent representations of DNNs, but extractable relations are mostly limited to semantic similarity. As a next step towards *QR for validation and verification of DNNs*: Concretely, we propose a method that extracts the learned superclass hierarchy from a multimodal DNN for a given set of leaf concepts. Under the hood we (1) obtain leaf concept embeddings using the DNN’s *textual input modality*; (2) apply hierarchical clustering to them, using that *DNNs encode semantic similarities via vector distances*; and (3) label the such-obtained parent concepts using search in *available ontologies from QR*. An initial evaluation study shows that meaningful ontological class hierarchies can be extracted from state-of-the-art foundation models. Furthermore, we demonstrate how to validate and verify a DNN’s learned representations against given ontologies. Lastly, we discuss potential future applications in the context of QR.

1 Introduction

One of the basic ingredients of QR models is an ontology specifying the allowed concepts, relations, and any prior assumption about them; more precisely, the commitment to (a subset of an) ontology with associated semantic meaning of concepts and relations [20]. Thanks to years of research, large and rich ontologies like Cyc [30], SUMO [35], or ConceptNet [53] are readily available for building or verifying QR models.

Meanwhile, however, DNNs have become the de-facto state of the art for many applications that hardly allow a precise input specification [42], such as processing of raw images (*computer vision*), e.g., for object detection [19], or processing of unstructured natural language text [37]. This machine learning approach owes its success to

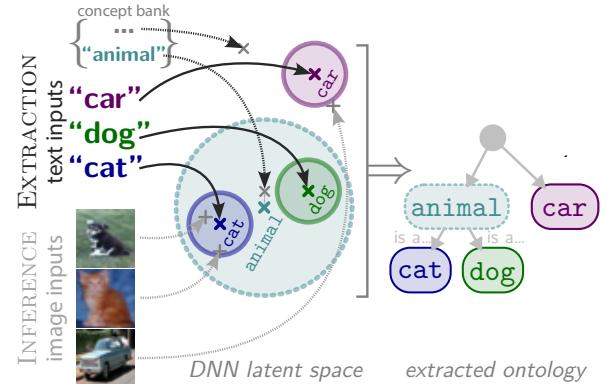


Figure 1. Illustration of the approach for ontology extraction from multi-modal DNNs: For extraction, (1) obtain leaf nodes (cat, dog, car) as the latent representations of their textual descriptions; (2) cluster these to get parent representations (dotted); (3) assign parents the closest concept (animal) from a concept bank. For inference check at each level similarity against nodes’ latent representations (e.g., first animal vs. car).

its strong representation learning capabilities: DNNs automatically learn highly non-linear mappings (*encoding*) from inputs to vectorial intermediate representations (*latent representations* or vectors) [11], and reasoning-alike processing rules [3] [23] from these to a desired output. Availability of large text and image datasets have further sparked the development of *multimodal* so-called *foundation models* [10] [28] [45]. These are large general-purpose DNNs trained to develop semantically rich encodings suitable for a variety of tasks [10]. This is oft achieved by training them to map textual descriptions and images onto matching vectorial representations (*text-to-image alignment*) [45], using multimodal inputs of both images and text.

The prospect. Foundation models come with some interesting prospects regarding their learned knowledge: (1) One can expect foundation models to **learn a possibly interesting and useful ontology**, giving insights into *concepts* [27] [29] [49] [62] and concept relations [16] [27] prevalent in the training data; and (2) such sufficiently large models can also **develop sophisticated reasoning chains** on the learned concepts [23] [44]. From the point of perspective of QR, this raises the question, whether this learned knowledge is consistent with the high quality available ontologies and QR models. This opens up well-grounded verification and validation criteria for safety or ethically critical applications. As a first step towards this, this pa-

* Corresponding Author. Email: mert.keser@continental.com

** Corresponding Author. Email: gesina.schwalbe@uni-luebeck.de

¹ Equal contribution.

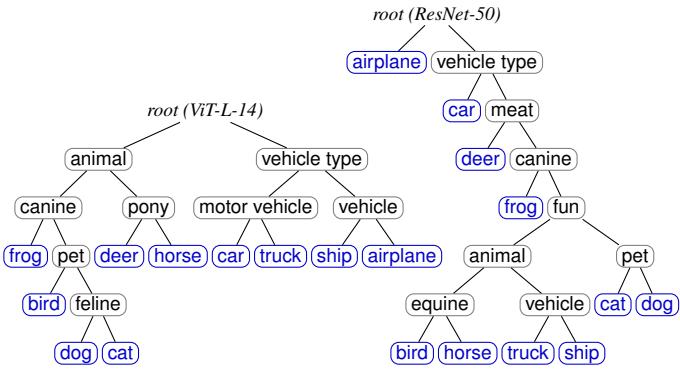


Figure 2. Comparison of two superclass hierarchies for given leaf concepts (blue) from CIFAR-10 [4] extracted from the large ViT-L-14 (left; with optimized prompt; 92% accuracy) and the smaller ResNet-50 (right; 46% accuracy) CLIP backbones with optimal distance metric settings. It shows the positive influence of model quality and prompt optimization (using “*a photo of a class*” instead of “*class*”) on the plausibility of the extracted ontology, and how the human-alignedness accuracy serves as indicator for it.

per defines techniques for extraction and verification of simple class hierarchies. Future prospects encompass to use the extracted knowledge from DNNs for knowledge retrieval, and ultimately gain control over the learned reasoning: This would enable the creation of powerful **hybrid systems** [14, 31] that unite learned encoding of raw inputs like images with QR models.

The problem. Unfortunately, the flexibility of DNNs in terms of knowledge representation comes at the cost of *interpretability* [22]; and, being purely statistical models, they may extract *unwanted and even unsafe correlations* [27, 47, 51]. The opaque distributed latent representations of the input do not readily reveal which interpretable concepts have been learned, nor what reasoning is applied to them for obtaining the output. This is a pity, not least because that hinders verification of ethical and safety properties. Take as an example the ontological commitment: Which hierarchical subclass-relations between concepts are considered? An example is shown in Fig. 3. This directly encodes the learned bias, which commonalities between classes are taken into account, and which of these are predominant for differentiating between classes. The same example also nicely illustrates the issue with wrongly learned knowledge: The models may focus on irrelevant but correlated features to solve a task, such as typical background of an object in object detection [47].

$$\begin{aligned} \text{(a)} \text{mammal} &\supseteq \{\text{cat}, \text{dog}, \text{horse}\}, \text{amphibian} \supseteq \{\text{frog}\} \\ \text{(b)} \text{indoor} &\supseteq \{\text{cat}, \text{dog}\}, \text{outdoor} \supseteq \{\text{horse}, \text{wet}\}, \text{wet} \supseteq \{\text{frog}\} \end{aligned}$$

Figure 3. Two exemplary ontological commitments: class hierarchies of the given leaf classes *frog*, *cat*, *dog*, *horse*, differentiating by (a) biology (mammal vs. amphibian), (b) image background (a Clever Hans effect!).

A whole research field, *explainable artificial intelligence* (XAI), has evolved that tries to overcome the lack of DNN interpretability [22, 50]. To date it is possible to partly associate learned representations with interpretable symbolic *concepts* (1-ary predicates) [52], such as whether an image region is a certain object part (e.g., *isLeg*), or of a certain texture (e.g., *isStriped*) [16, 27]. However, extraction of learned relations is so far focused on simple semantic similarity of concepts [16, 48]; hierarchical relations that hold across subsequent layers, i.e., across subsequent encoding steps [27, 59, 60]; or hierarchies obtained when subdividing a root concept [33]. And while first works recently pursued the idea to extract superclass hierarchies from given leaves, these are still limited to simple classifier architectures [59]. A next step must therefore be: Given a set of (hierarchy leaf) concepts, how to extract (1) the **unifying superclasses**, and (2) the

resulting **class hierarchy with subclass relationships** from any semantically rich intermediate output of a DNN, preferably from the embedding space of **foundation models**.

Approach. We here propose a simple yet effective means to get hold of these encoded class hierarchies in foundation models; thereby taking another step towards unveiling and verifying the ontological commitment of DNNs against known QR models respectively ontologies. Building on [59] and [62], our approach leverages two intrinsic properties of the considered computer vision models:

- (1) Vision DNNs generally encode learned concept similarities via distances in their latent representation vector space [16]. This makes it reasonable to find a hierarchy of superclass representations by means of **hierarchical clustering** [59].
- (2) Foundation models accept textual descriptions as inputs, trained for **text-to-image alignment**. This allows to cheaply establish an approximate bijection of textual concept descriptions to representations: A description is mapped by the DNN to a vector representation, and a given representation is assigned to that candidate textual description mapped to the most similar (=close by) vector [62].²

Contributions. Our main contributions and findings are:

- ★ An approach to **extract and complete a simple learned ontology**, namely a superclass hierarchy with given desired leaf concepts (Figure 2), from intermediate representations of any multimodal DNN, which allows to manually validate DNN-learned knowledge against QR models (see Figure 1);
- ★ An approach to **test the consistency of multimodal DNNs against a given class hierarchy**, e.g., from standard ontologies;
- ★ An initial experimental validation showing that the approach can **extract meaningful ontologies**, and reveal inconsistencies with given ontologies;
- ★ A thorough discussion of **potential applications** for QR extraction and insertion from / into DNNs.

2 Related Work

Extraction of learned ontologies. Within the field of XAI [22, 50], the subfield of concept-based XAI (c-XAI) has evolved around the goal to associate semantic concepts with vectors in the latent representations [29, 40, 49]. For analysis purposes, methods here allow to both extract representations which match given concept specifications (supervised approach) [16, 26, 27, 62] as well as mine meanings for the most prevalent representations used by the DNN (unsupervised approach) [18, 63]. Notably, we here utilize the supervised approach by Yuksekgonul et al. [62] which directly utilizes the text-to-image alignment in multimodal DNNs. Such associations have found manifold applications in the inspection of DNNs’ learned ontology, such as: Which concepts from a *given* ontology are learned [2, 52]? And how similar are representations of different concepts [16, 48]? This was extended to questions about the QR of the models, such as sensitivity of later concept representations (or outputs) to ones in earlier layers [27], or compliance with pre-defined logical rules [52]. However, very few approaches so far explored more *specific relations* between concept representations within *the same* layer’s representation space. In particular, specific relations beyond general semantic similarity, such as class hierarchies. This is a severe gap when

² This could be replaced by the mentioned approximate concept extraction techniques for models without decoder and text-to-image alignment.

trying to understand the learned ontological relations between concepts: DNNs develop increasing levels of abstraction across subsequent layers [16], rendering the concepts occurring in their representation spaces hardly comparable. Notably, Wan et al. [59] challenged this gap and applied hierarchical clustering on DNN representations. However, their association of given concepts to latent representations is limited to last layer’s output class representations, which we want to resolve. Furthermore, existing work was devoted only to single kinds of relations. We here want to show that these efforts can be unified under the perspective of investigating ontological commitment of DNNs.

3 Background

3.1 Deep neural network representations

DNNs. Mathematically speaking, deep neural networks are (almost everywhere) differentiable functions $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ which can be written in terms of small unit functions, the so-called *neurons* $f: \mathbb{R}^n \rightarrow \mathbb{R}$, by means of the standard concatenation operation $f \circ g: x \mapsto f(g(x))$, linear combination $x \mapsto Wx + b$, and product $a, b \mapsto a \cdot b$. Typically, the linear weights W and biases b serve as trainable parameters, which can be optimized in an iterative manner using, e.g., stochastic gradient descent. Neurons are typically arranged in *layers*, i.e., groups where no neuron receives outputs from the others. Due to this “Lego”-principle, DNNs are theoretically capable of approximating any continuous function (on a compact subspace) up to any desired accuracy [25], and layers can be processed highly parallel. In practice, this is a double-edged sword: DNNs of manageable size show astonishing approximation capabilities for target functions like detection or pixel-wise segmentation of objects in images [28, 56]. However, they also tend to easily extract irrelevant correlations in the data, leading to incorrect [47] or even non-robust [55] generalization respectively “reasoning” on new inputs.

Latent representations. In the course of an inference of an input x , each layer L of the DNN produces as intermediate output a vector $F_{\rightarrow L}(x) \in \mathbb{R}^n$, each entry being the output of one of the n neurons of L . This vectorial encoding of the input is called the *latent representation* of the input within L , and the vector space \mathbb{R}^n hosting the representations is called the *latent space*. Interestingly, it was shown that DNNs encode semantically meaningful information about the input in their latent representations, with abstraction increasing the more layers are passed (e.g., starting with colors and textures, to later develop notions of shapes and objects) [16, 36].

Concept embeddings. An emergent property of these representations is that in some layers, a concept C (e.g., color `Red`, or object part `Leg`), can be encoded as prototypical vector $e(C)$ within this latent space. These are called *concept (activation) vectors* [27] or *concept embeddings* [16]. The mapping $e: \mathcal{C} \rightarrow \mathbb{R}^n$ from a set of human-interpretable concepts to their embeddings even preserves semantic similarities to some extend: Examples are the reflection of analogical proportions [43] in word vector spaces (DNNs with textual inputs trained for natural language processing), like “ $e(\text{King}) - e(\text{Queen}) = e(\text{Man}) - e(\text{Woman})$ ” [32]; and their analogues in standard computer vision architectures trained for object classification or detection: “ $e(\text{Green}) + e(\text{Wood}) = e(\text{Tree})$ ” [16]. Our approach relies on these natural translation of semantic to vector operations/properties. In particular, we assume that the relation IsSimilarTo^3 on input instances x is mapped to some distance met-

³ We here assume that `IsSimilarTo` is reflexive and symmetric, following geometrical instead of psychological models of similarity [57].

ric d like Euclidean or cosine distance by the DNN representations: $\forall c, c': \text{IsSimilarTo}(c, c') \Leftrightarrow d(e(c), e(c')) \approx 0$ ⁴

Concretely, we use the translation of similarity relations to find a superclass concept representation via interpolation.

Text-to-image alignment. In the case of multimodal DNNs that accept both textual and image inputs, the training often encompasses an additional (soft) constraint: Given textual descriptions of an input image, these must be mapped to the same/a similar latent representation as their respective image. While pure language models suffer from the impossibility to learn the true meaning of language concepts without supervision [9], this additional supervision might help the model to develop representations that better match the human understanding of the word/concept. We here leverage this intrinsic mapping to associate textual or graphical descriptions of our concepts with latent representations.

When using textual descriptions, good text-to-image alignment is an important assumption; but, sadly, even with explicit training constraints this is not guaranteed [17] (cf. distance of image and text embeddings in Figure 4). We show both the influence of text-to-image alignment on our method, how it can be reduced, and how to use our method in order to identify issues with the learned meaning of concepts, which opens up options to fix the representations.

3.2 Ontologies

When modeling any problem or world, a basis of the model is to know “what the model is talking about”. This is exactly answered by the underlying *ontology*, i.e., a definition of what categories/properties and relations are used in the model. We here adopt the definition from [20].

Definition 1 (Ontology). *An ontology is a pair $(\mathcal{V}, \mathcal{A})$ constituted by a vocabulary $\mathcal{V} = \mathcal{C} \cup \mathcal{R}$ of a set of unary predicates \mathcal{C} (the concepts corresponding to class memberships and other properties) and a set of binary predicates \mathcal{R} (the instance relations) used to describe a certain reality, and which are further constraint by a set \mathcal{A} of explicit assumptions in the form of a first- (or higher-)order logic theory on the predicates.*

A relation we will use further is `IsSimilarTo` $\in \mathcal{R}$. Also spatial relations like `IsCloseBy` [52] and `LeftOf`, `TopOf`, etc. [44] have been defined and used in literature for latent space representations of objects. Simple examples of assumptions that relate the concept sets are, e.g., the subclass relationship we investigate in this paper: $\text{IsSuperclassOf}(c', c) : \Leftrightarrow (\forall v: c(v) \Rightarrow c'(v))$ (cf. Figure 3). This can also be seen as a relation between concepts, by interpreting the unary concept predicates C as sets of objects (e.g., classes) via $v \in C : \Leftrightarrow C(v)$. The validity of concept embeddings also gives rise to assumptions about concepts $(\forall v: C(v) \Leftrightarrow \text{IsSimilarTo}(v, e(C)))$. Note that, given embeddings, we can formulate relations between concepts using *instance* relations $R \in \mathcal{R}$ via $R(c, c') : \Leftrightarrow R(e(c), e(c'))$. An example would be `isSimilarTo(cat, dog)`.

The first challenge in extracting learned QR from DNNs is to find/explain the ontology that is used within the reasoning process of the DNN. Unraveling an ontology as done in [1] above breaks this step roughly down into:

- (1) Find the concepts \mathcal{C} (and their embeddings) used by the model.

⁴ For optimization, the relative formulation can be more convenient:
 $\forall c, c', c'': c \text{ more similar to } c' \text{ than to } c'' \Rightarrow d(e(c), e(c')) \leq d(e(c), e(c'')).$

- (2) Find the relations \mathcal{R} that may be formulated on vector instances.
- (3) Simple assumptions $\mathcal{A}_s \subseteq \mathcal{A}$: How are concept related.

(4) Identify further assumptions $\mathcal{A} \setminus \mathcal{A}_s$ that the model applies.

Note that the layer-wise architecture of DNNs partitions the representations into objects (vectors) in the different latent spaces. For a layer L we denote v in the latent space of L as $L(v)$. This gives rise to a partition of the concept, relation, and assumption definitions, allowing to conveniently split up above steps as follows:

- (1') What concepts $\mathcal{C}_i \subset \mathcal{C}$ are encoded *within the i th layer L_i* ($\forall C \in \mathcal{C}_i, v: \neg L_i(v) \Rightarrow \neg C(v)$)?

(3a') What assumptions $\mathcal{A}_{i,i}$ hold for which items within *the same i th latent space* ($\forall A \in \mathcal{A}_i, (v^{(s)})_s: \bigvee_s \neg L_i(v^{(s)}) \Rightarrow \neg A(v^{(1)}, \dots)$)?

(3b') What assumptions $\mathcal{A}_{i,j}, i \neq j$, hold between items of *different latent spaces*?

Task (1') is (somewhat) solved by methods from c-XAI, where both learned concepts [16] [27] [62] as well as their distribution over different layer representation spaces [34] are investigated. (3a') and (3b') show the yet-to-be-filled gaps: Investigated relations between items, item groups respectively concepts within the same arbitrary latent space (= (3a')). These so far only concern general semantic similarity, and relations across latent spaces only sensitivity. That falls far behind the richness of natural language; in particular it misses out on concept and instance relations of the kind “ c is similar to c' with respect to feature F ” respectively “ c, c' both are F ”, and counterpart “ c differs from c' with respect to feature F ⁵”. In other words, the relation `IsSuperclassOf` is missing, despite known to be learned [59]. This inhibits the expressivity of extracted constraints such as obtained in [44], as this directly relies on the richness of available vocabulary. The method proposed in this paper thus sorts in as follows: **We extend the extraction of relations relevant to point (3a')** (**relations amongst concepts within the same layer representation space**) **by allowing to extract the `IsSuperclassOf` relation between concepts**.

3.3 Hierarchical clustering

Hierarchical clustering [46] aims to find for a given set M a chain of partitions $\mathcal{M}_1 \leq \mathcal{M}_2 \leq \dots \leq \{M\}$ connected by inclusion⁶, i.e., assign each point in M to a chain of nested clusters $M_{1,i_1} \subseteq M_{2,i_2} \dots \subseteq M$, as illustrated in Figure 1. Such a hierarchy can be depicted using a dendrogram as in Figure 2. There are two regimes for hierarchical clustering: Divisive breaks up clusters top-down, while agglomerative starts from the leaves $\mathcal{M}_1 = \{\{p\} \mid p \in M\}$ and iteratively merges clusters bottom-up [46]. We here employ hierarchical clustering to find a hierarchy of subsets of latent representation vectors. Since we start with given leaf vectors, **this work uses standard agglomerative hierarchical clustering [61]**⁷. This optimizes the partitions for small distance between the *single points* within a cluster (*affinity*) and a large distance between the *sets of points* making up different clusters (*linkage*), typically at a complexity of $\mathcal{O}(|M|^3)$.

⁵ “ c, c' both are F ” ($\forall x: (c(x) \vee c'(x)) \Rightarrow F(x)$) rewrites to `IsSuperclassOf`(F, C) \wedge `IsSuperclassOf`(F, C'); the “differs”-case to \neg `IsSuperclassOf`(F, C) \wedge `IsSuperclassOf`(F, C').

⁶ To be precise: $\mathcal{M} \leq \mathcal{M}' \Leftrightarrow \forall M \in \mathcal{M}: \exists M' \in \mathcal{M}': M \subseteq M'$

⁷ We here use the scikit-learn implementation at <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html>

4 Approach

This section details our approach towards extracting a globally valid approximation of a DNN’s learned concept hierarchy, given the hierarchy’s desired leaf concepts. The goal is to allow manual validation or verification testing against existing ontologies from QR. Recall that this both requires a guided exploration of the learned concepts (*which parent classes did the model learn?*), as well as an exploration of the applicability of the superclass relation (*which superclasses/features are shared or different amongst given concepts?*). We will start in subsection 4.2 by detailing how to obtain the extracted class hierarchy (here simply referred to as *ontology*). This is followed by an excursion on how to conduct a kind of instance-based inference using the global taxonomy (subsection 4.2), which is then used in subsection 4.3 where we discuss techniques for validation and verification of DNN learned knowledge.

4.1 Extracting an ontology

Overview The steps to extract our desired ontology are (explained in detail further below): (1) obtain the **embeddings** $e(c_i)$, (2) apply **hierarchical clustering** to obtain superclass representations as superclass cluster centers, (3) **decode** the obtained superclass representations into a human-interpretable description.

Ingredients. We need as ingredients our trained **DNN** F , some concept encoder e (in our case defined using the DNN, see Step 1 below), the finite set $(c_i)_i = \mathcal{C}_{\text{leaf}}$ of **leaf concepts** for which we want to find parents classes, and the choice of **layer** L in which we search for them. Furthermore, to ensure human interpretability of the results, we constrain both our leaf concepts as well as our solution parent concepts to come from a given **concept bank** \mathcal{C} of human-interpretable concepts⁸. We furthermore need per concept $c \in \mathcal{C}$: A **textual description** `toText`(c) of c as textual specification; optionally a set `toImages`(c) containing the concept as graphical specification (see Step 1), as available, e.g., from many densely labeled image datasets [8] [24]; and optionally a set `Parents`(c) of candidates for parent concepts of c (for more efficient search). The following assumptions must be fulfilled, in order to make our approach applicable:

Assumptions 1.

- (a) **Text-to-image alignment:** The DNN should accept textual inputs, and be trained for text-to-image alignment, such that for a suitable textual description T of any concept $c \in \mathcal{C}$ one can reasonably assume $e(c) \approx F_{\rightarrow L}(T)$. We use this to find embeddings: The embedding of a visual concept c can be set to the DNN’s text encoding $F_{\rightarrow L}(T)$ of a suitable textual description T of c .
- (b) **Existence of embeddings:** For all leaf concepts, embeddings $e(c_i)$ of sufficient quality exist in the latent space of L .
- (c) **Concentric distribution of subconcepts:** Representations of subconcepts are distributed in a **concentric manner** around its parent. Generally, this does not hold [33], but so far turned out to be a viable simplification as long as semantic similarities are well preserved by the concept embedding function e [18] [41]. I.e. for a superclass concept `Parent` with children set \mathcal{C}_S we can choose

$$e(\text{Parent}) \approx \underset{\text{Child} \in \mathcal{C}_S}{\text{mean}} e(\text{Child}) \quad (1)$$

- (d) **Semantic interpolability:** Consider a latent representation v that is close to or inbetween (wrt. linear interpolation) some embeddings $e(C_i)$ and $e(C_j)$. We assume that v can be interpreted

⁸ The concept bank restriction makes this essentially a search problem.

to correspond to some concept, i.e., $\exists c \in \mathcal{C} : \|e(c) - v\|_2 < \epsilon$ for some admissible error ϵ . This is needed to make the averaging in the parent identification in (I) above meaningful.

Note that Assumption I(d) is very strong, stating that there is a correspondence between the semantic relations of natural language concepts, and the metric space structure of latent spaces. This is by no means guaranteed, but according to findings in word vector spaces [32] and also image model latent spaces [16] a viable assumption for the structure of learned semantics in DNNs.

Step 1: Obtain the embeddings $e(c_i)$. We here leverage the text-to-image alignment to directly define the concept-to-vector mapping $e: e(c) := \text{mean}_{x \in \text{toDNNInput}(c)} F_{\rightarrow L}(x)$. Following [59, 62], the `toDNNInput` function can be a mapping from concept to a single textual description [62] or to a set of representative images [59].

- **Textual concepts:** The naive candidate for a textual description $\text{toDNNInput}(c) := \text{toText}(c)$. However, some additional prompt engineering may be necessary, i.e., manual adjustment and finetuning of the formulation [17, 45]. For example, following [45] we replace “ c ” by “an image of c ” for the prompting.
- **Visual concepts:** Here we take the graphical `toImages(c)` specification of our concept. One could then employ standard supervised c-XAI techniques to find a common representing vector for the given images, e.g., as the weights of a linear classifier of the concept’s presence [16, 27]. We here instead simply feed the DNN with each of the images and capture its respective intermediate latent representations, which is valid due to the concentricity assumption.

If the text-to-image alignment is low, we found image representations of concepts to yield more meaningful results.

Step 2: Hierarchical clustering. Employ any standard hierarchical agglomerative clustering technique to find a hierarchy of partitions of the set of given concept embeddings. Each partitioning level represents one level of superclasses, with one cluster per class (see the simple example in Figure 1). As of (I), the mean of the cluster’s embedding vectors is the embedding of its corresponding superclass (*the cluster center*).

Note that the hierarchical clustering in principle allows to: (a) start off with more than one vector per leaf concept, e.g., coming from several image representations or from jointly using embeddings from textual and image representations; (b) weight the contribution of each child to the parent. This, however, is only viable together with means to automatically determine the weights, and not further pursued here.

Step 3: Decoding of cluster centers. We here use a two-step search approach to assign each cluster center a concept from the concept bank \mathcal{C} . Given a cluster center p , the first optional step is to reduce the search space by selecting a subset of candidate concepts from \mathcal{C} . Following [62], (1a) we collect for every leaf concept c the set of those concepts that, according to the ConceptNet knowledge graph [53], are related to c by any of the relations in $\mathcal{R}_{\text{concepts}} = \{\text{hasA}, \text{isA}, \text{partOf}, \text{HasProperty}, \text{MadeOf}\}$:

$$\text{Parents}(c) := \{p \mid \bigvee_{R \in \mathcal{R}_{\text{concepts}}} R(p, c)\}. \quad (2)$$

(1b) The union $\mathcal{P} = \bigcup_{c \text{ leaf in cluster}} \text{Parents}(c)$ of these sets serves as candidate set for p . Note that this is a simplification that allows to capture as superclass any best fitting commonality between the leaf concepts (e.g., background context like `indoor` or biological relation like `mammal` for `{cat, dog}` as in Figure 3). Generally, there is a trade-off between very specific relation definitions, and fidelity to

the learned knowledge of the model. The trade-off can be controlled by the broadening or narrowing of the candidate set. The here chosen broad definition of the `IsSuperClass` relationship between concepts favors fidelity to the model’s learned knowledge. Investigating effects of more narrow concept candidate sets is future work. (2) In the second step, the concept for p is then selected from the candidate set \mathcal{P} to be the one with minimum distance embedding (embeddings again obtained as in Step 1): $e^{-1}(p) := \operatorname{argmin}_{p \in \mathcal{P}} \|p - e(p)\|_2$.

The final result then is a hierarchy tree, where leaf nodes are the originally provided concepts, inner nodes are the newly extracted superclasses, and the connections represent the `IsSuperclassOf` relation. In the experimental section we will more closely investigate the influence of the proposed variants with/without prompt engineering and with/without finetuning.

4.2 Inference of an ontology

The such obtained ontology can be used for outlier-aware inference, i.e., classification of new input samples to one of the leaf concepts. This will be useful not only as an interesting standalone application in safety-relevant classification scenarios, but in particular for the validation.

The baseline of the inference is the k -nearest neighbor classifier: It directly compares the latent representation of a new input with each available concept embedding; and then assigns the majority vote of the k nearest concept embeddings. To enrich the inference process with information from the ontology, one instead traverses the ontology tree, at each node branching off towards the closest child node.

Remark 1. Note that this allows to easily insert an outlier criterion: If at a parent class p none of the children nodes is closer than a threshold, the sample is considered an outlier of class p . This neatly preserves the maximum amount of information available about the properties of the sample, and, thus, eases subsequent handling of the unknown input. For example, an outlier of (parent-)class `StaticObject` should be treated differently than one of (parent-)class `Animal`.

Hyperparameters of this inference procedure are the choice of similarity, including whether to take into account the size (variance/width) of the cluster, e.g., by favoring wide over near-to-point-estimate clusters; and the threshold for being an outlier.

4.3 Validating and comparing learned ontologies

We now get to the core goal of this paper: Verify or validate a given DNN using QR. For this we start with validation of an extracted ontology from subsection 4.1 and discuss how to measure its fidelity to DNN learned knowledge, and alignedness to human prior knowledge, which here corresponds to the expected image-to-concept matching. Lastly, we show how one can encode a given ontology as contextualized embeddings to verify a DNN against given prior knowledge from QR.

Human-alignedness. One main desirable of a DNN’s ontology is that it well aligns with the semantics that humans would expect and apply for the respective task. Any mismatch may either bring insights to the human on alternative solutions, or, more probably, indicates a suboptimal solution or even Clever Hans effect of the learned representations. A straight-forward way to measure the human-alignedness is to test the **prediction accuracy** of the ontology when used for inference (see subsection 4.2) on human-labeled samples. If human labels deviate often from the predictions, this indicates a bad alignment of the semantics the DNN has learned for the

concepts from those a human would expect. Other means to estimate the human-alignedness (not yet investigated in this work) are direct qualitative user studies, where human evaluators **manually check** the consistency of the obtained ontology tree with their own mental model; or automatic checking of consistency against given world knowledge or common sense ontologies like Cyc [30]. Lastly, the improvement in humans' predictions about the behavior of the model, a typical human-grounded XAI metric [50], could quantify in how far humans can make sense of the ontology.

A different aspect of human-alignedness is how well the ontology, in particular the inference scheme it defines, generalizes to novel concepts (semantic outliers) that so far have not occurred in leaves or nodes. The generalization can be measured as the performance in assigning a correct parent node. A special case here are blended cases where the novel concept unifies features of very different classes, such as a *cat* with *wheel* as walking support. The uncertainty of the model in such blended cases can be qualitatively compared against human one, potentially uncovering a bias.

Text-to-image alignment. The to-be-expected performance of cross-modal inference of the ontology (i.e., ontology defined using textual concepts, but inference done on images) directly depends on the quality of the text-to-image alignment. This motivates a use as an indicator for suboptimal text-to-image alignment.

Fidelity. Fidelity of the ontology, respectively shortcomings in the simplified modeling of the ontology, can be measured by the deviation between the baseline inference directly on the leaves, and the ontology inference. Inference on the leaf concepts c_i means we predict for an image x the output class c for which the textual embedding is closest to the embedding of x , proximity measured with respect to some distance d (here: cosine similarity):

$$c := \operatorname{argmin}_{c \in (c_i)_i} d(F_{\rightarrow L}(\text{toText}(c), F_{\rightarrow L}(x))) \quad (3)$$

This is referred to as naive *zero-shot* approach, following research on using foundation models on specialized tasks without finetuning (=with training on zero samples) [17, 45]. The reason to choose this as a baseline is that the ideal tree should sort samples into the same leaf neighborhood as direct distance measurement would do. Simplifications that may infringe this equality are unequal covariances (\approx widths) of sibling class clusters; the chosen similarity measure; or assuming perfect text-to-image alignment.

Verification against a given ontology. The previous extraction techniques yield an inspectable representation of the ontology learned by a model. This allows manual validation of the learned knowledge against models from QR. Alternatively, one could directly verify a multimodal model against consistency with a given ontology: In short, we propose to modify the leaf concept embeddings from Step 1 such that they additionally encode their local part of the ontology, i.e., information about all desired parents of the leaf, as *context*. One can then measure the performance of naive inference (see subsection 4.2) on these contextualized leaf nodes as defined in (3). A higher performance then means a better alignment of the context of a leaf concept with its image representations. This even would allow to narrow down unalignedness to specific concepts (those with bad inference results). We suggest as point of attack for contextualization is the textual encoding: Let c be a leaf concept at depth d in the tree with chain of parents $(p_i)_{i=1}^d$ from root to leaf. We can now follow [17] and modify the original $\text{tT} = \text{toText}$ function of a leaf concept to:

$$\text{toText}'(c) := "\text{tT}(p_1), \dots, \text{tT}(p_d), \text{tT}(c)" \quad (4)$$

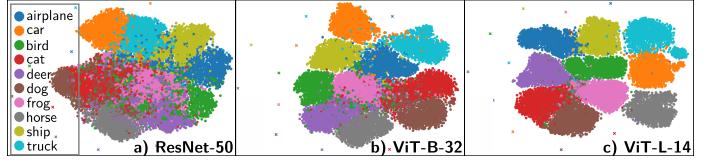


Figure 4. Visualization of the latent space representations of CIFAR-10 embeddings in different CLIP model backbones (one color per class), generated using the distance-preserving t-SNE dimensionality reduction method [58]. The better class separation in the transformer-based backbones (b, c)) are consistent with fidelity and human-alignedness results in Tabs. 1 2

E.g., *cat* may turn into “animal, pet, cat”. The effect is that the obtained embedding (possibly after prompt finetuning as above) is shifted towards including the desired context; and all leaves together encode the complete ontology.

5 Experiments

5.1 Settings

Models under test. In our experiments, we utilized CLIP [45], one of the first multimodal foundation model family accepting both text and images [13]. For text-to-image alignment CLIP was trained to map an image and its corresponding text descriptions onto a similar (with respect to cosine similarity) latent space representation. This general-purpose model captures rich semantic information, and achieves impressive performance compared to task-specific models across various applications, including image captioning [7, 12], recognition of novel unseen objects [5], and retrieval tasks [6, 54]. This makes it a common choice as basis for training or distilling more specialized models [12, 13], and thus a highly interesting target for validation and verification of its learned knowledge and internalized QR. In our experiments, we explored various CLIP backbones, including ResNet-50, as well as Vision Transformer (ViT) variants featuring different patch sizes and model capacities (e.g., ViT-B/32, ViT-L/14)⁹

Dataset. The CIFAR-10 dataset [4] Chap. 3] is a benchmark in the field of computer vision, consisting of 60,000 32×32 color images, split into 50,000 training and 10,000 test images. The images are equally distributed onto the 10 diverse classes *airplane*, *ship*, *car*, *truck*, *bird*, *cat*, *dog*, *deer*, *horse*, *frog*. The choice of classes suits our initial study well, as they both exhibit pairs of semantically similar objects (e.g., *car*, *truck*), as well as mostly unrelated ones (e.g., *car*, *cat*), so we can expect a deep class hierarchy. In our study, we conduct inference both of the baseline (naive zero-shot) and the proposed method on the CIFAR-10 test dataset [4].

Fidelity baseline. As discussed in subsection 4.3, the inference on the leaf concepts (naive zero-shot approach) serves as baseline (maximum performance) for fidelity measurements. The closer the tree inference gets to the naive zero-shot performance, the higher the fidelity. We here choose as distance metric the cosine distance $\text{CosDist}(a, b) := 1 - \frac{a \cdot b}{\|a\| \cdot \|b\|}$ (0 for a, b parallel, 1 for orthogonal, 2 for $a = -b$), going along with the training of CLIP.

Metrics. Any quantitative classification performances are measured in terms of **accuracy** of the results on CIFAR-10 test images against their respective ground truth label.

⁹ Pre-trained models and weights were obtained from: <https://github.com/openai/CLIP>

5.2 Ablation Study: Influences on Human-Alignedness and Fidelity of Ontology Extraction

As detailed in subsection 4.3, to measure the **human-alignedness** of the given multi-modal encoder model, we evaluated the performance when using our extracted ontology for inference of class labels on new images. And as a **fidelity indicator**, we measure the performance drop between inference on the leaves (naive zero-shot approach) against that of inference on our tree.¹⁰ Both are measured in the course of an ablation study to identify the influence of different settings on the ontology’s usefulness and quality.

Investigated influences. Both the ontology extraction by means of agglomerative hierarchical clustering (see subsection 3.3) as well as later the inference on new samples (see subsection 4.2) rely on measuring similarities between embedding vectors. However, due to being automatically optimized, the embeddings’ optimal similarity metric is unknown. Hence, we treat each choice of similarity metric as a hyperparameter, and investigate their influence on human-alignedness of the extracted ontology:

- **Affinity:** Affinity typically influences which data points are most similar, i.e., closest related, in the final tree structure. In our experiments, we tested the standard Manhattan (L_1), Euclidean (L_2), and cosine distances.
- **Linkage:** This parameter determines the criterion used to merge clusters during the hierarchical clustering process, and in particular affects the shape and compactness of the clusters. In our experiments, we tested the standard settings of Ward, complete, average, and single linkages. Ward linkage minimizes the variance within clusters, while complete / average / single linkage focuses on the maximum / average / minimum distance between clusters.
- **Inference similarity:** We use the same choices as for affinity. Next, we compare different settings for obtaining the leaf embeddings. The following variants are considered:
 - **Prompt tuning:** In case text embeddings are to be obtained, CLIP suggests using text prompts in the form “*a photo of a classname*” rather than simply “*classname*”, because the model is trained on image captions as text. If applied, this augmentation is done for both leaf and parent node textual embeddings.
 - **Text encoding vs. few-shot image encoding:** As described in subsection 4.1 Step 1, the two different approaches to obtain leaf embeddings are text encoding and image encoding. We here only consider few-shot image encoding, i.e., specifying the concept via < 10 images, which ensures manageable complexity of the hierarchical clustering algorithm.¹¹

Results. An illustrative example of an ontology extracted from CLIP (ViT-L-14 backbone) using the prompt “*a photo of a classname*” is provided in Figure 2 for found-to-be-optimal settings according to the ablation study. Consistently optimal hyperparameter settings with respect to human-alignedness and fidelity turned out to be affinity=Manhattan, linkage=complete, and inference similarity=cosine, which were also used to create the remainder of the ablation studies. The accuracy results on CIFAR-10 of inference using the extracted ontology versus the naive-zero shot approach as a baseline for fidelity are given in Tabs. 1 for the prompt engineering settings, and 2 for the comparison of text and image encodings of the leafs.

Please note that we did not yet conduct a cross-validation, so results should foremostly serve as guide for further investigations.

¹⁰ Performance against a ground truth is only a proxy; future experiments should directly compare predictions of the two.

¹¹ Standard implementations have a complexity of $\mathcal{O}(n^3)$ for n leaf samples.

Table 1. Comparison of inference accuracy using naive zero-shot (Naive) and our method across different model architectures and textual prompt types. Fidelity calculated as ratio $\frac{\text{ours}}{\text{naive}} \in [0, 1]$; best models marked.

	Prompts					
	“classname”			“a photo of a classname”		
	Naive	Ours	ratio	Naive	Ours	ratio
ResNet-50	0.70	0.46	0.66	0.69	0.67	0.97
ViT-B-32	0.87	0.82	0.94	0.89	0.85	0.96
ViT-L-14	0.91	0.85	0.93	0.95	0.92	0.97

First findings. In advance we manually validated the assumption of a good text-to-image alignment (Assumption I(a)). For this we visualized the distribution and class separability of text and CIFAR-10 test sample embeddings in the latent spaces of the different CLIP backbones, results shown in Figure 4. The dimensionality-reduced visualizations suggest that with increasing parameter number, the clusters of different classes become more distinctly separated; and transformer-based backbones demonstrate superior separation. Notably, across all backbones, the text inputs and images are encoded in separate regions of the latent space, indicating a clear distinction between these two modalities in the model’s internal representation.

The *prompt engineering*, i.e., replacing the text prompt “*classname*” with “*a photo of classname*” turned out to be have a strong positive impact on human-alignedness and fidelity in case of the worse aligned CNN-based CLIP backbone, and still a notable one for the already good transformer backbones.

In contrast, using *few images instead of text to obtain the leaf embedding resulted in worse performance*. However, in our initial tests performance seemed to increase with the number of images: Dropping the few-shot constraint showed competitive results. In the following table, we replaced the leaf node information with the randomly-sampled training images in the respective class.

Table 2. Comparison of inference accuracy for different ways to obtain the leaf embeddings: *few-shot* image embeddings vs. textual embeddings (*zero-shot*), with the naive zero-shot approach as baseline. Best model **bold**.

	Few-Shot			Zero-Shot	
	1-shot	5-shot	10-shot	Naive	Ours
ResNet-50	0.45	0.58	0.61	0.69	0.67
ViT-B-32	0.67	0.79	0.86	0.89	0.85
ViT-L-14	0.64	0.76	0.80	0.95	0.92

It should be noted, that a better performance of the textual embedding could possibly be attributed to a sub-optimal text-to-image alignment. This would be consistent with the insights into the distribution and class separability of image and text embeddings in the latent space in Figure 4 (with respect to Euclidean distance). It should be further investigated, whether this must be attributed to disparity in metrics, the domain shift to CIFAR-10 inputs, or could serve as an indicator for bad text-to-image alignment wrt. the considered classes.

5.3 Ontology validation and verification

Validation: qualitative results. A manual inspection of the obtained ontologies (see Figure 2 for an example) showed, that *good human-alignedness also coincides with seemingly valid tree structures*. Seemingly valid here means, that a human inspector can easily find convincing arguments for the validity most of the splitting criteria of the nodes. In Figure 2 two trees which are created with different parameters are compared. The tree on the left, which uses ViT-L/14 as a backbone, affinity clustering, and Manhattan linkage, achieves 92% accuracy on the classification task. In contrast, the tree

on the right, created with a ResNet-50 backbone, affinity clustering, and Euclidean linkage, yields an accuracy of 45%. One of the reasons for the low accuracy score in the classification task for the tree on the right is that its decision process does not align well with human-like decision-making. For example, the structure first checks whether an object is a “*vehicle*” and then whether it is “*meat*”. This decision process deviates from human-aligned reasoning, which can also be observed through manual inspection.

Furthermore, we identified the tendency that the *superior vision transformer backbones also showed the seemingly more valid tree structures*. This possible architectural dependency of good ontological commitment should be further investigated.

Verification against a given ontology. To exemplify the verification of ontological commitment against a given ontology, we chose the simple tree structure provided by [59] for CIFAR-10 dataset. To label the inner nodes of this tree, we utilized two external knowledge sources: WordNet [15] and GPT-4 [1], in each case bottom-to-top queried for a textual description of a parent for sibling nodes. We then used the ontology information to create contextualized leaf embeddings, as described in subsection 4.3 and applied naive zero-shot inference on these contextualized leaves. For WordNet, we labeled each node with the closest matching superclass. For GPT-4, we queried the model to provide the superclass of the given leaf nodes.

Initial verification results for the different given ontologies are shown in Table 3. As expected, using the extracted learned ontology for the contextualization caused no change compared to the baseline of non-contextualized embeddings; this contextualization is supposed to be equivalent to the non-contextualized leaf embeddings from the perspective of the model. However, the contextualization with external ontologies caused a strong drop in inference accuracy. A closer look at the results showed that those leaves with parents mentioning technical terms (e.g., “non-mammalian vertebrate”) were mostly misclassified, indicating that the learned knowledge is inconsistent / not aware of these parts of the given ontologies. Further research is needed on practical implications (e.g., thus induced error cases), and how to align the ontologies.

Table 3. Verification results of different models against different sources of external ontologies: the NBDT tree structure [59] with WordNet [15] or GPT-4 [1] queried node labels; versus no contextualization (*Naive*) and contextualization against the extracted ontology (*Ours*). Values are measured in inference accuracy on contextualized nodes.

	WordNet	GPT-4	Naive	Ours
ResNet-50	0.31	0.36	0.69	0.67
ViT-B-32	0.40	0.53	0.89	0.85
ViT-L-14	0.52	0.54	0.95	0.92

6 Future work: Applications and next steps

6.1 Applications of learned ontology extraction

Our method opens up several further interesting applications for the use of QR in DNN understanding, verification, and improvement.

Optimal learned reasoning representations. As discussed above, access to the internal ontology of a DNN is key to understand its internal QR. In particular, an open research question is, *what kind of concept representations are DNNs optimized for*, and, subsequently, *which kinds of reasoning would be supported by this?* For example, qualitative spatial reasoning would most benefit from a region-based representation of concepts, while cone-based reasoning from cones

as representations [38]. The quantitative measurement of ontological commitment allows to do ablation studies on different representations of concepts and relations, e.g., different similarity measures.

DNN inspection. The obtained ontologies open up new inspection possibilities for DNNs. An interesting one could be to generate **contrastive examples** [21]: Change a given input minimally such that the class/superclass changes, possibly under a constraint to remain within a given superclass. Also, one could globally test the models against biases towards scenarios respectively background. A bias is uncovered, if the commonality of two classes is based on background rather than functionally relevant features; possibly supported on test samples generated by inpainting techniques. Unfortunately, the text-to-image alignment training of foundation models may easily introduce such a bias, as concepts occurring in similar image scenarios additionally will occur in similar textual context. E.g., one may expect *cat* and *dog* to be similar, as both often occur indoors.

Knowledge insertion. The final goal of the introspection discussed above should be to not only be able to verify the learned ontological commitment, but also to control both the commitment, and subsequently the learned reasoning. This might be achieved by adding penalties during training, determined by iterative ontology extraction and model finetuning. Thus, a foundation model with acceptable ontological commitment may be obtained. Lastly, to distill this knowledge of the large model into smaller specialized models, standard model distillation techniques could be amended [39]. Concretely, regularization terms can be added to (1) enforce that correspondences to some/most of the concepts, and to (2) enforce respective similarities and other relationships between the concepts.

6.2 Next steps

Our initial experiments are clearly limited in their extend, so immediate next steps should encompass more experiments on measuring **human-alignedness** respectively a larger ablation study on possible influence of the made assumptions. Such can be domain shifts, like text-to-image, and real-to-synthetic image. Experiments should include user studies, and comparison to existing ontologies; Similarly, the **outlier detection and handling** capabilities of ontologies should be further investigated, both for novel as well as novel blended classes. Lastly, it can be investigated how to extend the here proposed approach **from multimodal models to unimodal** ones, allowing to compare the ontologies of large foundation models against that of state-of-practice small and efficient object detectors.

7 Conclusion

Altogether, this paper tackles the problem how to validate and verify a multimodal DNN’s learned knowledge using QR. Concretely, we take the step to unveil the ontological commitment of DNNs, i.e., the learned concepts and (here: superclass-)relations. For this, we proposed a simple yet effective approach to (1) uncover yet undiscovered superclasses of given subclasses as used by the DNN; and to (2) extract a full hierarchical class tree with the `IsSuperClass`-relationships; together with means to verify and validate the extracted part of the learned ontology. Even though this initial proof-of-concept still relies on some simplifications, our initial experiments could already extract meaningful class hierarchies from concurrent multimodal DNNs, and reveal inconsistencies with existing ontologies. These may serve as a basis to access further insights into the ontological commitment of DNNs, and subsequently validate and

verify its learned QR. We are confident that, eventually, this could allow to control, i.e., correct and integrate, valuable prior knowledge from QR into DNNs, creating powerful yet verifiable and efficient hybrid systems. Thus, we hope to spark further interest into interdisciplinary research of QR for verification of DNNs within the QR community.

Acknowledgements

The paper was written in the context of the "NXT GEN AI Methods" research project funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK). The authors would like to thank the consortium for the successful cooperation.

References

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] A. Agafonov and A. Ponomarev. An Experiment on Localization of Ontology Concepts in Deep Convolutional Neural Networks. In *Proc. 11th Int. Symposium on Information and Communication Technology, SoICT '22*, pages 82–87. Assoc. for Computing Machinery, 2022. doi: 10.1145/3568562.3568602.
- [3] F. Aghaeipoor, M. Sabokrou, and A. Fernández. Fuzzy rule-based explainer systems for deep neural networks: From local explainability to global understanding. *IEEE Transactions on Fuzzy Systems*, 2023.
- [4] K. Alex. Learning Multiple Layers of Features from Tiny Images. Master's thesis, University of Toronto, Canada, Apr. 2009.
- [5] N. Amini-Naieni, K. Amini-Naieni, T. Han, and A. Zisserman. Open-world text-specified object counting. In *BMVC*, 2023.
- [6] A. Baldriati, M. Bertini, T. Uricchio, and A. Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *Proc. IEEE/CVF Conf. CVPR*, pages 21466–21474, 2022.
- [7] M. Barraco, M. Cornia, S. Cascianelli, L. Baraldi, and R. Cucchiara. The unreasonable effectiveness of clip features for image captioning: an experimental analysis. In *Proc. IEEE/CVF Conf. CVPR*, 2022.
- [8] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proc. 2017 IEEE Conf. CVPR*, pages 3319–3327. IEEE Comput. Society, 2017. doi: 10.1109/CVPR.2017.354.
- [9] E. M. Bender and A. Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proc. 58th Ann. Meeting ACL*, pages 5185–5198. ACL, 2020. doi: 10.18653/v1/2020.acl-main.463.
- [10] R. Bommasani and et al. On the Opportunities and Risks of Foundation Models, 2022.
- [11] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proc. IEEE/CVF Int. Conf. Comput. vision*, pages 9650–9660, 2021.
- [12] J. Cho, S. Yoon, A. Kale, F. Dernoncourt, T. Bui, and M. Bansal. Fine-grained Image Captioning with CLIP Reward. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 517–527. Association for Computational Linguistics, July 2022. doi: 10.18653/v1/2022.findings-naacl.39.
- [13] CMA. AI Foundation Models: Initial Report. Technical report, Competition & Markets Authority, UK, Sept. 2023.
- [14] I. Donadello, L. Serafini, and A. S. d'Avila Garcez. Logic tensor networks for semantic image interpretation. In *Proc. 26th Int. Joint Conf. Artificial Intelligence*, pages 1596–1602. ijcai.org, 2017. doi: 10.24963/ijcai.2017/221.
- [15] C. Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010.
- [16] R. Fong and A. Vedaldi. Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *Proc. 2018 IEEE Conf. CVPR*, pages 8730–8738. IEEE Comput. Society, 2018. doi: 10.1109/CVPR.2018.00910.
- [17] Y. Ge and et al. Improving Zero-Shot Generalization and Robustness of Multi-Modal Models. In *Proc. IEEE/CVF Conf. CVPR*, pages 11093–11101, 2023.
- [18] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems 32*, pages 9273–9282, 2019.
- [19] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of field robotics*, 37(3):362–386, 2020.
- [20] N. Guarino. Formal Ontologies and Information Systems. In *Proc. FOIS'98*, pages 3–15. IOS Press, June 1998.
- [21] R. Guidotti. Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery*, 2022. doi: 10.1007/s10618-022-00831-6.
- [22] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang. XAI—Explainable artificial intelligence. *Science Robotics*, 4(37), 2019. doi: 10.1126/scirobotics.aa7120.
- [23] C. He, M. Ma, and P. Wang. Extract interpretability-accuracy balanced rules from artificial neural networks: A review. *Neurocomputing*, 387: 346–358, 2020.
- [24] J. He, S. Yang, S. Yang, A. Kortylewski, X. Yuan, J.-N. Chen, S. Liu, C. Yang, Q. Yu, and A. Yuille. PartImageNet: A Large, High-Quality Dataset of Parts. In *ECCV 2022*, pages 128–145. Springer Nature Switzerland, 2022. doi: 10.1007/978-3-031-20074-8_8.
- [25] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991. doi: 10.1016/0893-6080(91)90009-T.
- [26] M. Keser, G. Schwalbe, A. Nowzad, and A. Knoll. Interpretable model-agnostic plausibility verification for 2d object detectors using domain-invariant concept bottleneck models. In *Proc. IEEE/CVF Conf. CVPR*, pages 3890–3899, 2023.
- [27] B. Kim, M. Wattberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proc. 35th Int. Conf. Machine Learning*, volume 80 of *Proc. Machine Learning Research*, pages 2668–2677. PMLR, 2018.
- [28] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick. Segment Anything. In *Proc. IEEE/CVF Int. Conf. on Comput. Vision*, pages 4015–4026, 2023.
- [29] J. H. Lee, S. Lanza, and S. Wermter. From Neural Activations to Concepts: A Survey on Explaining Concepts in Neural Networks, 2024.
- [30] D. B. Lenat. *Building Large Knowledge-Based Systems*. Addison-Wesley Pub. Co., 1989. ISBN 978-0-201-51752-1.
- [31] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *Int. Conf. Learning Representations*, 2018.
- [32] T. Mikolov, W.-t. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proc. 2013 Conf. North American Chapter ACL: Human Language Technologies*, pages 746–751. ACL, 2013.
- [33] G. Mikriukov, G. Schwalbe, C. Hellert, and K. Bade. GCPV: Guided Concept Projection Vectors for the Explainable Inspection of CNN Feature Spaces, 2023.
- [34] G. Mikriukov, G. Schwalbe, C. Hellert, and K. Bade. Revealing similar semantics inside CNNs: An interpretable concept-based comparison of feature spaces. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Springer, 2023.
- [35] I. Niles and A. Pease. Towards a standard upper ontology. In *Proceedings of the International Conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, pages 2–9. Assoc. for Computing Machinery, 2001. doi: 10.1145/505168.505170.
- [36] C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. doi: 10.23919/distill.00007.
- [37] D. W. Otter, J. R. Medina, and J. K. Kalita. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624, 2021. doi: 10.1109/TNNLS.2020.2979670.
- [38] Ö. L. Özcep, M. Leemhuis, and D. Wolter. Embedding Ontologies in the Description Logic ALC by Axis-Aligned Cones. *JAIR*, 78:217–267, 2023. doi: 10.1613/jair.1.13939.
- [39] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016. doi: 10.1109/SP.2016.41.
- [40] E. Poeta, G. Ciravegna, E. Pastor, T. Cerquitelli, and E. Baralis. Concept-based Explainable Artificial Intelligence: A Survey, 2023.
- [41] A. F. Posada-Moreno, N. Surya, and S. Trimpe. ECLAD: Extracting Concepts with Local Aggregated Descriptors. *Pattern Recognition*, 147: 110146, 2024. doi: 10.1016/j.patcog.2023.110146.
- [42] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, and S. S. Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 51(5):1–36, 2018.

- [43] H. Prade and G. Richard. Analogical proportions: Why they are useful in AI. In *Proc. 30th Int. Joint Conf. Artificial Intelligence, IJCAI 2021*, pages 4568–4576. ijcai.org, 2021. doi: 10.24963/IJCAI.2021/621.
- [44] J. Rabold, G. Schwalbe, and U. Schmid. Expressive explanations of DNNs by combining concept analysis with ILP. In *KI 2020: Advances in Artificial Intelligence*, Lecture Notes in Comput. Science, pages 148–162. Springer, 2020. doi: 10.1007/978-3-030-58285-2_11.
- [45] A. Radford and et al. Learning Transferable Visual Models From Natural Language Supervision. In *Proc. 38th Int. Conf. on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [46] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu. Comprehensive survey on hierarchical clustering algorithms and the recent developments. *AI Review*, 56(8):8219–8264, 2023. doi: 10.1007/s10462-022-10366-3.
- [47] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, KDD '16*, pages 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778.
- [48] G. Schwalbe. Verification of size invariance in DNN activations using concept embeddings. In *Artificial Intelligence Applications and Innovations*, IFIP Advances in Information and Communication Technology, pages 374–386. Springer, 2021. doi: 10.1007/978-3-030-79150-6_30.
- [49] G. Schwalbe. Concept Embedding Analysis: A Review. *arXiv:2203.13909 [cs, stat]*, 2022.
- [50] G. Schwalbe and B. Finzel. A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 2023. doi: 10.1007/s10618-022-00867-8.
- [51] G. Schwalbe, B. Knie, T. Sämann, T. Dobberphul, L. Gauerhof, S. Raafatnia, and V. Rocco. Structuring the safety argumentation for deep neural network based perception in automotive applications. In *Computer Safety, Reliability, and Security. SAFECOMP 2020 Workshops*, pages 383–394. Springer, 2020. doi: 10.1007/978-3-030-55583-2_29.
- [52] G. Schwalbe, C. Wirth, and U. Schmid. Enabling verification of deep neural networks in perception tasks using fuzzy logic and concept embeddings, 2022.
- [53] R. Speer, J. Chin, and C. Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *Proc. AAAI Conf. Artificial Intelligence*, 31(1), 2017. doi: 10.1609/aaai.v31i1.11164.
- [54] M. Sultan, L. Jacobs, A. Stylianou, and R. Pless. Exploring clip for real world, text-based image retrieval. In *2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2023.
- [55] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [56] M. Tan, R. Pang, and Q. V. Le. EfficientDet: Scalable and efficient object detection. In *Proc. 2020 IEEE/CVF Conf. CVPR*, pages 10781–10790, 2020.
- [57] A. Tversky. Features of similarity, 1977. ISSN 1939-1471.
- [58] L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.
- [59] A. Wan, L. Dunlap, D. Ho, J. Yin, S. Lee, S. Petryk, S. A. Bargal, and J. E. Gonzalez. NBDT: Neural-backed decision tree. In *Posters 2021 Int. Conf. Learning Representations*, 2020.
- [60] D. Wang, X. Cui, and Z. J. Wang. CHAIN: Concept-harmonized hierarchical inference interpretation of deep convolutional neural networks. *CoRR*, abs/2002.01660, 2020.
- [61] J. H. Ward Jr. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi: 10.1080/01621459.1963.10500845.
- [62] M. Yukselgonul, M. Wang, and J. Zou. Post-hoc Concept Bottleneck Models. In *ICLR 2022 Workshop on PAIR2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022. doi: 10.48550/arXiv.2205.15480.
- [63] R. Zhang, P. Madumal, T. Miller, K. A. Ehinger, and B. I. P. Rubenstein. Invertible concept-based explanations for CNN models with non-negative concept activation vectors. In *Proc. 35th AAAI Conf. Artificial Intelligence*, volume 35, pages 11682–11690. AAAI Press, 2021.

Preliminary Experiments of Qualitative Reasoning Model Construction Using Large Language Model

Shinpei Suzuki^{a,*}¹ and Masaharu Yoshioka^{a, b,1}

^aGraduate School of Information Science and Technology, Hokkaido University

^bFaculty of Information Science and Technology, Hokkaido University

ORCID (Shinpei Suzuki): <https://orcid.org/0009-0001-7599-1617>, ORCID (Masaharu Yoshioka): <https://orcid.org/0000-0002-2096-1218>

Abstract. Qualitative Reasoning (QR) is a reasoning framework for simulating physical behavior based on naive knowledge of the physical system. However, it is not easy to build a model of the physical system based on physical laws and principles. Therefore, example-based model libraries are used to support the creation of a model. This approach requires extensive knowledge to represent a variety of physical systems. If suitable components are not prepared for a problem, new components must be added to the library.

Therefore, the main focus of this work was on developing a method to utilize Large Language Model (LLM) to solve this problem. Although LLM alone doesn't have the architecture necessary to perform QR, it performs well at discovering information such as physical phenomena and objects related to the problem if a sophisticated prompt can be prepared. To this end, we proposed a new model construction method using LLM as a tool to extract the fragmentary information. This information is used as a key to access the previously prepared database to get the physical parameter relations based on physical laws and principles. We introduce and validate this framework using a simple motion example that considers both spring motion and friction.

1 Introduction

Qualitative Reasoning (QR) is a reasoning framework for simulating physical behavior based on naive knowledge about the physical system. Qualitative Process Theory (QPT) [2] is one of the methods that uses knowledge about qualitative relationships between physical parameters. This method is good at representing the physical system using generalized concepts such as physical phenomena [4] and anchor concepts [3]. This framework can simulate the physical system well if the user succeeds in creating an appropriate model for the system. However, although the system provides a basic vocabulary to construct a model of the physical system, it is not easy to construct a model using these vocabularies. Therefore, the utilization of example-based model libraries such as physical feature(PF) [4] and subclass of anchor concepts [3] has been proposed in the literature. This approach requires a good amount of knowledge to represent varieties of physical systems.

Recently, Large Language Models (LLMs) such as ChatGPT² are

used in various tasks. LLMs are trained on a large variety of documents and can predict physical behavior based on the trained knowledge. However, the quality of the result of this prediction is not good enough because LLMs are trained only on the textual contexts and generate statistically plausible text, not correct reasoning [3]. However, based on the preliminary analysis of LLM and considering that they are trained on a large number of texts containing descriptions of a variety of physical systems, they can be used as a retriever to make a list of the related physical laws and principles for the given situations. Therefore, a new framework for qualitative reasoning model construction using LLM was proposed here. In this framework, the description of the physical system is provided as text, and LLM helps to collect general model fragments and relationships among them to support the model construction process. By using this framework, we assumed that the size of the knowledge for describing varieties of physical systems used in the previous methods can be reduced. We introduce and validate this framework using a simple motion example that considers spring motion and friction.

2 Model Construction for Qualitative Reasoning

Although the knowledge used for qualitative reasoning is general and reusable, it is necessary to construct a model for behavioral simulation. Therefore, Kiriyama et al. [4] proposed to use physical features(PFs) that represent typical physical systems with related physical phenomena. The Knowledge Intensive Engineering Framework [7] supports the construction of a model for qualitative reasoning by combining these building blocks. Nonetheless, it is necessary to create example-based libraries to represent the variety of components used in the physical system. Anchor concepts [3] also have a similar problem for model construction. For example, the "motion" has 355 subclasses to represent different situations. It is desirable to have a general framework for constructing a model based on an understanding of the physical system configuration.

Recently, it has been demonstrated that LLMs can perform well in various tasks. So LLMs' ability for QR has been experienced (e.g. for design [5] and spatial reasoning [1]). As Forbus says [3], their success criterion is the generation of statistically plausible text, not correct reasoning. On the other hand, LLM can entail possible physical phenomena that occur on the given physical system configuration (e.g., sliding entails the possibility of friction). This feature was used here to support the QR model construction.

* Corresponding Author. Email: suzuki.shinpei.h8@elms.hokudai.ac.jp

¹ Equal contribution.

² <https://www.openai.com/chatgpt> (Last accessed on June 19, 2024.)

Table 1 summarizes the comparison of Kiriyma, Forbus, and our method. Primitive concepts are atomic representation of physical behavior.

3 A New Model Construction Method

LLMs perform well at gathering information about cases that are similar to a given physical system configuration. However, their ability to combine this information varies. For example, they excel at tasks such as gathering and then comparing information about two subjects, namely A and B, a process that can be expressed using a chain of thought (CoT) [6]. However, they are not necessarily good at solving problems that require synthesizing multiple pieces of knowledge based on first principles, such as estimating behavior based on physical knowledge. To address this issue, this study exploits the LLM's ability to skillfully gather case-related information by collecting data related to objects, processes, and physical parameters. This information is then used in a framework that employs a QR model to infer behavior. The output of the LLM can also be viewed as information that solves the problem of creating case-specific PFs needed to describe when certain physical phenomena occur.

This method exploits the strengths and compensates for the weaknesses of LLMs. In this research, we used ChatGPT (GPT-4) to solve a problem involving the motion of a single particle on a frictional constraint surface, without dealing with specific numerical values.

To construct the proposed method, we created a database for QR with physical laws and principles (Figure 1). This database stores the physical laws and principles corresponding to each process, and further contains physical parameters related to these laws and principles, along with their qualitative temporal changes and constraints. By extracting the process names from the output of LLM, the associated physical laws and principles are compiled with their corresponding physical parameters and relationships, instantiated on a per-object basis. In this step, due to the variations of description about physical laws and principles generated by LLM, we manually rewrite the description for finding out the data stored in the database.

This instantiation details the relationships between the generic physical parameters defined in the process for the specific object, thereby creating a physical parameter network used to construct the model.

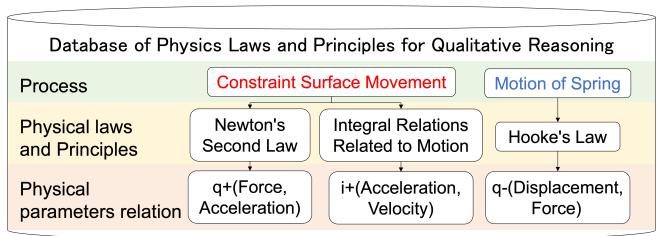


Figure 1. Database of Physics Laws and Principles for Qualitative Reasoning

The proposed method suggests a three-step approach for building the model and inferring behavior (Figure 2).

- (1) **Analysis of Physical Problems** Using LLM to extract necessary information such as objects, processes, and missing physical parameters from the text of physical problems.
- (2) **Construction of a Physical Parameter Network** Using the process names as keys to access corresponding physical laws and principles from a database of physical laws and principles for

QR, instantiate the qualitative time-varying relationships between the associated physical parameters for each object, and construct a physical parameter network.

- (3) **Calculation of System State Transitions** Obtain the initial values of the physical parameters from the LLM. And based on these initial values and the qualitative time-varying relationships between the physical parameters, calculate the possible state transitions.

The following sections explain this method with a practical example. The problem addressed is "The spring attached to the wall was pulled sufficiently in the opposite direction of the wall and then released along the rough floor". See Appendix A for details on the prompt and output used for ChatGPT. Information for the next step are manually extracted from the output.

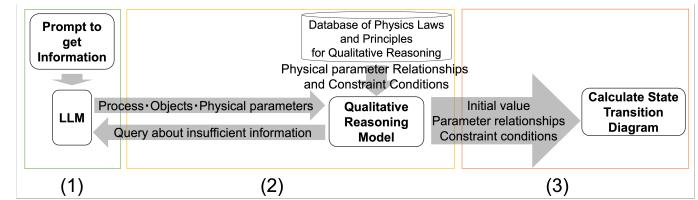


Figure 2. Proposed framework of Qualitative Reasoning

3.1 Analysis of a Physical Problem

We enter the text about a physical problem with an instruction prompt to LLM for identification of physical system configuration (objects and their contact states) and a list of physical phenomena that occurred in the system. This instruction prompt is designed for extracting such information based on the flow described in Figure 3.

The details of the procedure are as follows. First, the objects appearing in the problem are recognized and it is determined whether their motion conforms to a constraint surface, such as a floor. If the motion is along a constraint surface, the axis and slope are determined by checking whether the constraint surface is a slope, for example. Next, it detects how the objects are in contact with the constraint surface, whether vertically or horizontally. Then it extracts the information corresponding to the process. With these steps, a prompt template for analyzing the basic information of the problem is created and entered into ChatGPT along with the problem being handled. From the example input, the output was able to extract objects such as a spring and a wall, and a horizontal floor as a constraint surface. However, the wall is considered immobile and is not treated as an object. Similarly, processes such as Constraint surface movement, Friction movement, and Spring movement were applicable as inferred from the content obtained.

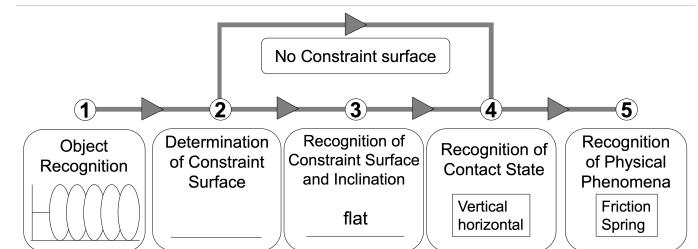


Figure 3. Procedure of analyzing problem

Table 1. Comparison of the Physical feature, Anchor concepts, and the Proposed method

	Kiriyama et. al., 1992 [4]	Forbus 2023 [3]	proposed method(Ours)
Primitive Concepts	Physical Phenomena	Anchor concepts	Physical Phenomena
Example Based Components	Physical feature	Subclass of Anchor concepts	
Difficulty of Model Construction	Easy	Easy	Supported by LLM
Size of Knowledge Base	Large	Large	Small

3.2 Construction of the Physical Parameter Network

A physical parameter network is constructed from the information obtained.

3.2.1 Generation of Relationships Between the Physical Parameters

Using the process names obtained from the analysis of the physics problem as keys, access is made to a pre-created database of physical laws and principles for QR (Figure 1). Based on the physical laws and principles associated with the processes, the qualitative time-varying relationships of the physical parameters are instantiated for each object.

In addition, forces are automatically aggregated by component along each axis. If information is missing, such as the conditions under which static friction occurs, ChatGPT is queried, and from the output, the action of static friction is defined. In this case, it corresponds to the restoring force of the spring.

Thus, for the example problem, a physical parameter network has been created concerning the qualitative time-varying relationships of instantiated physical parameters (Figure 4). However, due to the complexity, the constraints are omitted.

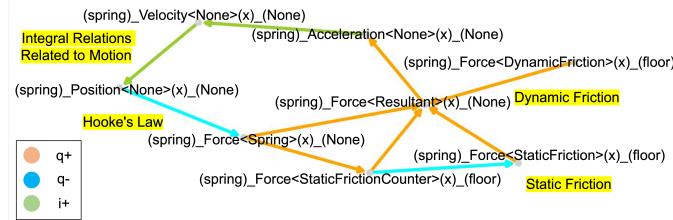


Figure 4. Physical parameter network

3.3 Calculation of System State Transitions

This section explains the method used to determine the system's state transitions by setting the initial values of the physical parameters in addition to the information from the physical parameter network created through the above processes. In this study, the values of the physical parameters are qualitative and indicate one of three directions: +, 0, or -.

3.3.1 Obtaining Initial Value Information

To get the initial value information of the physical parameters, we design CoT prompt for extracting initial value. In this prompt, we ask LLM to provide the information about the axis of motion, the origin, and the definition of positive and negative directions, first. Then, using CoT, we re-identify the problem, identify the axis and origin, positive and negative directions, and list the initial values for position and velocity. Next, we can list the acting forces. If friction is present, distinguish between static and dynamic friction and select the appropriate one, then re-list the acting forces, and list the resultant force and acceleration. A prompt containing these instructions is

typed into ChatGPT to obtain initial values, and the output was obtained. From this output, the initial values of the physical parameters for the example problem were set.

For this example text "The spring attached to the wall was pulled sufficiently in the opposite direction of the wall and then released along the rough floor", the system extract the initial values as follows; Position (+), Velocity (0), Acceleration (-), Resultant Force(-), Dynamic Friction (0), Static Friction (+), Reaction of Static Friction (-), Spring Force (-).

3.3.2 Envisioning

qlabelsec:envisionning Based on the obtained physical parameters, their qualitative temporal changes and initial values, an envisioning simulation of state transitions through QR is performed by a QPT-based simulator.

In the simulation results (Figure 5), each node in the diagram represents the state of the system at each point in time (Detail of the initial state and each states are shown in Appendix B). The orange node on the right represents a positive position, the yellow-green node in the middle represents the origin (the natural length of the spring), and the light blue node on the left represents a negative position. The edges represent the direction of the state transitions, and the state represented by the central red label indicates the final state. The resulting state transitions are mainly counterclockwise including periodic behavior of simple harmonic motion and transitions to a rest state due to friction.

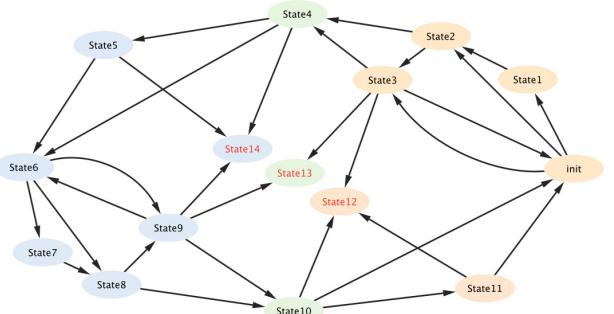


Figure 5. State Transition Diagram

4 Conclusion

In this research, we proposed a framework for the construction of a QR model using physical phenomena and relationships between the objects collected from text provided by the LLM. This model integrates first principles such as physical laws and principles to simulate the behavior. The operation of the framework was demonstrated using simple harmonic motion on a friction plane as an example. In the future, various challenges should be addressed including extending the knowledge of the physical laws and principles and applying the system to problems involving other physical phenomena. In addition, there is a need to build an end-to-end framework and provide feedback information on the state transitions of the inference results to

the LLM, which could extend the range of potential applications to generating explanations.

References

- [1] A. G. Cohn. An Evaluation of ChatGPT-4's Qualitative Spatial Reasoning Capabilities in RCC-8. In "Working papers, 36th International Workshop on Qualitative Reasoning", 2023. URL <https://staff.fnwi.uva.nl/b.bredeweg/QR2023/pdf/07Cohn.pdf>
- [2] K. D. Forbus. Qualitative Process Theory. *Artificial Intelligence*, 24: 85–168, 1984.
- [3] K. D. Forbus. Building Domain Theories for Commonsense Reasoning from LanguageGrounded Ontologies. In *Working papers, The 36th International Workshop on Qualitative Reasoning*, 2023. URL <https://staff.fnwi.uva.nl/b.bredeweg/QR2023/pdf/14Forbus.pdf>
- [4] T. Kiriyama, T. Tomiyama, and H. Yoshikawa. Building a Physical Feature Databasefor Integrated Modeling in Design. In *Working papers, The Sixth International Workshop on Qualitative Reasoning about Physical System*, pages 124–138, 1992.
- [5] J. T. Maxwell III, J. de Kleer, I. Matei, and M. Zhenirovsky. Preliminary Experiments using LLMs for Design. In "Working papers, 36th International Workshop on Qualitative Reasoning", 2023. URL <https://staff.fnwi.uva.nl/b.bredeweg/QR2023/pdf/13Maxwell.pdf>
- [6] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b.ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf
- [7] M. Yoshioka, Y. Umeda, H. Takeda, Y. Shimomura, Y. Nomaguchi, and T. Tomiyama. Physical Concept Ontology for the Knowledge Intensive Engineering Framework. *Advanced Engineering Informatics*, 18(2):95–113, 2004.

A Prompt and output of the system

A.1 Prompt1

System message

Please consider the following "movement" within the scope of high school physics.

Please identify the following materials related to the movement of the following "movement".

Please answer in the form of an itemized list for each of the following materials.

Please consider the Earth's gravity.

Please do not output any text other than the requested materials. **System message end**

Movement

The spring attached to the wall, stretched sufficiently, is about to start moving along the rough floor in the opposite direction of the wall, having been released.

Movement end

Materials

- **objects:** List all objects featured in the movement.
- **is constraint surface movement:** Please determine whether it is a motion along the constraint surface.
- **constraint surface:** List names and angles (flat, slope upward, slope downward, vertical etc..) of the constraint surface, if it exists.
- **constraint surface objects:** Extract objects from "constraint surface" that are not part of objects in "objects".
- **objects and constraint surface objects:** Combine "objects" and "constraint surface objects" into a bullet-point list.

Materials end

A.2 Output1

Materials

- **objects:**
 - Spring
 - Wall
- **is constraint surface movement:** Yes
- **constraint surface:**
 - Floor (flat)
- **constraint surface objects:**
 - Floor
- **objects and constraint surface objects:**
 - Spring
 - Wall
 - Floor

Materials end

A.3 Prompt2

Explain concisely what kind of motion will occur based on physical laws and principles for each object.

Let's think step by step about this logically.

A.4 Output2

When the spring, attached to the wall and stretched on a rough floor, is released, the following motion occurs based on physical laws and principles:

- **Elastic Potential Energy to Kinetic Energy:** The stretched spring holds elastic potential energy. Upon release, this energy converts into kinetic energy, causing the spring to move.
- **Motion Against Friction:** The spring moves on a rough floor, indicating the presence of friction. This friction opposes the motion of the spring, gradually reducing its kinetic energy.
- **Oscillatory Motion:** The spring will likely undergo oscillatory motion (back and forth) due to its elastic nature, until all its kinetic energy is dissipated through friction and internal damping within the spring.
- **Effect of Gravity:** Earth's gravity acts vertically downwards, but since the motion is horizontal on a flat surface, gravity primarily affects the normal force between the spring and the floor, which in turn influences the frictional force.

Overall, the motion involves a conversion of elastic potential energy to kinetic energy, oscillatory movement, and energy dissipation due to friction and internal damping, within the constraints of gravitational and frictional forces.

A.5 Create Physical parameter Network

The following can be interpreted from these outputs.

- **objects:**
 - Spring
- **constraint surface:**

- Floor (flat)

- **Physical phenomena:**

- Constraint surface movement
- Spring movement
- Friction

Then, access the database and create a physical parameter network (Figure: 4). However, since the restoring force of the spring is insufficient this time, inquire with the LLM.

A.5.1 Prompt

What force acts as the reaction to static friction?

A.5.2 Output

The reaction to static friction is the force exerted by an object against the surface it is in contact with. In the context of the spring attached to the wall and moving on a rough floor, when static friction is acting to prevent the spring from starting its motion, the reaction to this static friction is the elastic force (restoring force) exerted by the stretched spring in the direction away from the wall. This force is trying to move the spring across the floor, and the static friction is what resists this motion. According to Newton's third law of motion, for every action, there is an equal and opposite reaction. Thus, the force the spring exerts against the floor to move forward is matched by an equal and opposite force from the static friction preventing that motion.

A.5.3 Add information

Therefore, set the static friction force as the restoring force of the spring.

A.6 Prompt3

Direction

- **origin:** Point at the natural length of the spring
- **positive direction:** Away from the wall from the origin.
- **negative direction:** Toward the wall from the origin.

Direction end

List the initial signs of the physical parameters based on the "Direction".

1. **Step 1:** Check the "Movement".
2. **Step 2:** Check the "Direction".
3. **Step 3:** List the initial values of Position, Velocity.
4. **Step 4:** List all added forces for each object.
5. **Step 5:** If friction is applied, distinguish static friction and dynamic friction, and reconsider which is applied.
6. **Step 6:** List the initial values of all added Forces.
7. **Step 7:** List the initial values of resultant forces and Acceleration.

A.7 Output3

Based on the "Movement" and the defined "Direction", let's list the initial signs of the physical parameters:

- **Step 1: Movement** - A spring attached to a wall, stretched sufficiently, and about to start moving along the rough floor in the opposite direction of the wall, having been released.
- **Step 2: Direction** -
 - Positive Direction: Away from the wall.
 - Negative Direction: Toward the wall.
- **Step 3: Initial Values of Position and Velocity** -
 - Position: Positive (since the spring is stretched away from the wall).
 - Velocity: Zero initially (at the moment of release, the spring is momentarily stationary).
- **Step 4: Added Forces for Each Object** -
 - Spring: Elastic force (restoring force), frictional force (due to the rough floor), and gravitational force.
 - Wall: Reaction force exerted by the spring.
 - Floor: Normal force exerted by the spring.
- **Step 5: Friction** - Initially, static friction applies (as the spring is just about to move). Upon movement, it transitions to dynamic (kinetic) friction.
- **Step 6: Initial Values of all Added Forces** -
 - Elastic force: Negative (restoring force towards the wall).
 - Static friction: Zero initially (since the spring is not yet moving), but it will be positive (opposing the motion away from the wall).
 - Gravitational force: Neutral in terms of direction (acts downwards, perpendicular to the direction of spring's motion).
 - Normal force: Neutral in terms of direction (acts perpendicular to the motion).
- **Step 7: Initial Values of Resultant Forces and Acceleration** -
 - Resultant force on the spring: Initially negative (due to the restoring elastic force).
 - Acceleration: Initially negative (the spring accelerates towards the wall due to the restoring force).

B States generated by the envisioning system

Followings are description of the states generated by the envisioning system.

Init The length of the spring is longer than the normal length and it starts accelerating towards the wall (acceleration = -, velocity = 0, position = +).

State1 The length of the spring is longer than the normal length and it starts accelerating towards the wall and it starts moving towards the wall (acceleration = -, velocity = -, position = +).

State2 The acceleration is 0 because the force of the spring and the dynamic friction force are balanced (acceleration = 0, velocity = -, position = +).

State3 The acceleration direction changes because the dynamic friction force is greater than the spring force (acceleration = +, velocity = -, position = +).

State4 The length of the spring becomes the normal length (acceleration = +, velocity = -, position = 0).

State5 The length of the spring is shorter than the normal length and moves towards the wall (acceleration = +, velocity = -, position = -).

State6 The length of the spring is shorter than normal and stops (acceleration = +, velocity = 0, position = -).

State7 The length of the spring is shorter than normal and is moving away from the wall (acceleration = +, velocity = +, position = -).

State8 The acceleration is 0 because the spring force and the dynamic friction force are balanced (acceleration = 0, velocity = +, position = -).

State9 The acceleration direction changes because the dynamic friction force is greater than the spring force (acceleration = -, velocity = +, position = -).

State10 The length of the spring becomes the normal length (acceleration = -, velocity = +, position = 0).

State11 The length of the spring is longer than the normal length and moves away from the wall (acceleration = -, velocity = +, position = +).

State12 Length of spring is longer than normal length and stops (spring force and static friction force are balanced)

State13 Length of spring is normal length and stops (spring force and static friction force are balanced)

State14 Length of spring is shorter than normal length and stops (spring force and static friction force are balanced)

Using Qualitative Techniques with Kolmogorov-Arnold Networks for Explainable AI

Ismael Sanz^{a,*}, Lledó Museros^{a,1}, Vicente Casales-García^{b,1} and Luis González-Abril^{b,1}

^aUniversitat Jaume I, Spain

^bUniversidad de Sevilla, Spain

ORCID (Ismael Sanz): <https://orcid.org/0000-0001-9670-5627>, ORCID (Lledó Museros): <https://orcid.org/0000-0001-5521-2666>, ORCID (Vicente Casales-García): <https://orcid.org/0000-0001-8537-7023>, ORCID (Luis González-Abril): <https://orcid.org/0000-0002-2532-0946>

Abstract. Kolmogorov-Arnold networks (KANs) are neural networks that work by fitting a composition of simple univariate functions. They present several advantages with respect to perceptrons; in particular, they are capable of learning fully symbolic equations, thus generating inherently interpretable models. However, these symbolic representations are not generally easily human-understandable. Through a simple use case, we show how we can use qualitative techniques to find intuitive explanations for KAN-learned models. We show how KANs and qualitative techniques are complementary, and propose future avenues of research.

1 Introduction

Kolmogorov-Arnold Networks (KANs) [7] have recently emerged as a hot topic in the field of neural networks. KANs are based in the Kolmogorov-Arnold representation theorem [5], that states that any continuous multivariate function can be represented as a finite sum of continuous univariate functions and their compositions. Leveraging this theoretical foundation, Kolmogorov-Arnold Networks aim to decompose complex, high-dimensional functions into more manageable univariate components, thereby enhancing both interpretability and computational efficiency.

In the current landscape of Artificial Intelligence techniques, this approach promises several important advantages: firstly, KANs can be more parameter-efficient than an equivalent multi-layer perceptron. KANs are particularly useful in applications where the relationship between input variables is intricate and nonlinear. By breaking down these relationships into simpler, univariate functions, KANs can effectively capture the underlying patterns with fewer parameters, reducing the risk of overfitting. Furthermore, the modular nature of KANs allows for easier adaptation and extension, making them suitable for a wide range of tasks from regression and classification to more complex domains such as time-series prediction [9] and image processing [2].

Secondly, they can produce explainable results, since in principle it's possible to fit functions with a symbolic interpretation. This has motivated a flurry of applications where KANs are used to learn processes which can be modeled as relatively straightforward physics-informed equations. From our point of view this is particularly im-

portant, since traditional neural network architectures, while powerful, often operate as black boxes. In response, the field of *Explainable Artificial Intelligence* (XAI) [8] has emerged to study how to open these black boxes, which has important technical, ethical and legal implications. KANs are potentially very useful tools in that regard, as an approach which is both technically robust and interpretable in principle.

However, the fact that KANs are able to generate symbolic functions does not mean that these functions are readily human-interpretable. For that reason, we seek to introduce qualitative reasoning approaches into the KAN framework. Qualitative reasoning focuses on understanding and modeling the behavior of systems without relying solely on quantitative data, providing a complementary perspective that emphasizes the relationships and constraints inherent in the data. By combining KANs with qualitative reasoning, we can develop models that not only perform well but also provide deeper insights into the underlying mechanisms of the phenomena being studied.

The main goal of this paper is to present ideas on how KANs and qualitative techniques can be applied together, with a focus on Explainable AI. We will use a case study in which we use KANs to learn a simple color transformation in images, and we will use a qualitative theory to provide an intuitive explanation of the result, which complements the symbolic formula learned by the KAN. We will then discuss possible avenues of research for further integration of KANs and qualitative approaches.

The paper is structured as follows. Section 2 briefly introduces the motivating case study and how it is solved by using KANs. Section 3 shows how we can use qualitative techniques — in particular, the Qualitative Color Description (QCD) theory — to provide an intuitive explanation of the model learned by the KAN. Finally, Section 4 discusses potential areas of research, and the paper ends with some brief conclusions.

2 KANs: Motivating example

2.1 Brief introduction to KANs

The Kolmogorov-Arnold representation theorem [5] states that if f is a multivariate continuous function on a bounded domain, then it can be written as a finite composition of continuous functions of a single

* Corresponding Author. Email: isanz@uji.es

¹ Equal contribution.

variable using addition. More formally, for a smooth $f : [0, 1]^n \rightarrow \mathbb{R}$

$$f(x) = f(x_1, \dots, x_n) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^n \phi_{q,p}(x_p) \right)$$

, where $\phi_{q,p} : [0, 1] \rightarrow \mathbb{R}$ and $\Phi_q : \mathbb{R} \rightarrow \mathbb{R}$.

Several of these compositions can be combined as layers, thus creating a *Kolmogorov-Arnold Network* (KAN) of arbitrary depths and widths. In each edge of a KAN, there is a univariate function that is fitted to the training data. Thus, a suitable family of basis functions must be selected; the original implementation uses B-splines, but other options are certainly possible. In the next section we show how KANs work with an example.

2.2 Motivating example: Reconstructing watermarked images with a KAN

To illustrate the use of KANs, we'll use a very simple digital image processing example. Consider the process of *watermarking* images using a mask. For example, Figure 1 shows a sample photograph, and Figure 2 a watermarked version using a simple mask. Our task is to learn the transformation between the masked pixels in the original photograph and the corresponding ones on the watermarked image.



Figure 1. Original image

This transformation is better defined in a color model such as HSV (Hue-Saturation-Value), which uses human-understandable concepts rather than uninterpretable RGB values. Thus, after transforming the image to HSV, we learn three separate KANs, where each takes as input three separate values (H, S, V) and outputs the transformed Hue, Saturation and Value respectively. We try to keep the KANs as simple as possible, using the smallest network that provides a good result. Other than that, we do not perform any kind of hyperparameter optimization on the KANs. The implementation is done in Python using the PyKAN package.²

The learned KAN for Hue is shown in Figure 3. It's a three-layer KAN with three inputs and one output, and the fitted component functions are displayed.

PyKAN uses B-splines as basis functions. After the initial splines are fitted, it's possible to fit a well-known symbolic function that is approximated by these splines. For instance, the bottom-left function



Figure 2. Watermarked image

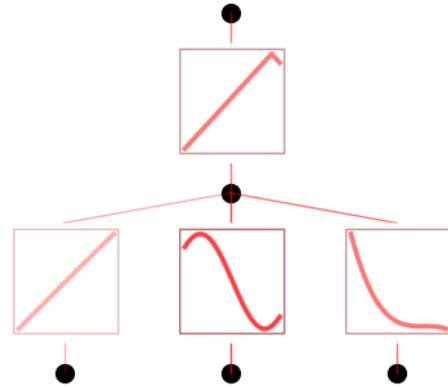


Figure 3. Trained KAN for Hue, showing the learned B-spline functions

can be interpreted as a polynomial, and the function to the left as a trigonometric function such as sine or cosine. The best-fitting function, if any, is selected from a set of well-known candidate functions. After this step, the KAN is retrained with the new component functions. This allows the creation of a fully symbolic representation for the result. In this case, the resulting function is

$$0.9 - 0.17 | -0.07 (0.77 - x_V)^3 - 0.04 \sin (4.63 x_S + 0.81) - 23.09 \tanh (0.25 x_H - 0.05) + 4.15 |$$

where x_H , x_S and x_V represent the input hue, saturation and value respectively. Note that, while symbolic, this can hardly be considered to be a human-friendly explanation of the model, even though it is arguably better in that respect than just having the weights of a neural network. To be fair, in this case it should be certainly possible to achieve a simpler expression by using more training data and a bit of hyperparameter tuning, but our point is that the fact that KANs can produce a symbolic expression does not automatically mean that the result is readily understandable by humans. Incidentally, the expression obtained for the Value KAN is slightly simpler, while the one for the Saturation is far more complex.

Figures 4 and 5 show the learned KANs for Saturation and Value respectively. The learning metrics are reasonable: from about 60000 pixels in the training set, we achieve test RMSE values of between 0.01 and 0.05, which are good enough for this simple case. The KANs train in a few minutes on a Macbook without special GPU support.

Thus, with these three KANs, we are able to fully specify the trans-

² <https://github.com/KindXiaoming/pykan>

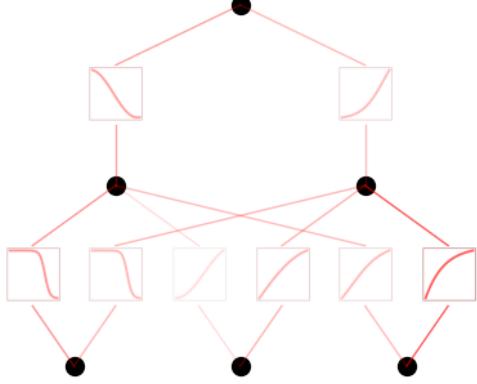


Figure 4. Trained KAN for Saturation

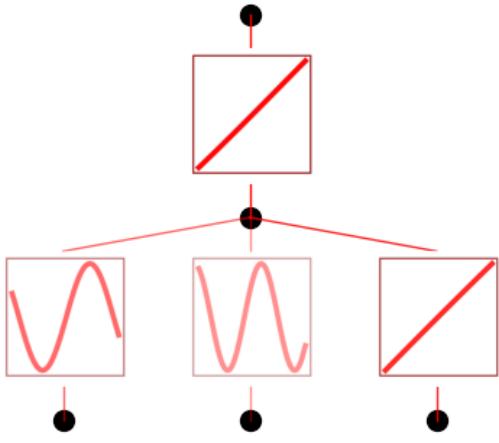


Figure 5. Trained KAN for Value

formation from the color of a pixel in the original image to its corresponding one in the watermarked version. However, it's not clear at all that this result is explainable to humans. In the following section we show how we can use a qualitative color model to better understand this result.

3 Improving interpretability with a qualitative model

3.1 The QCD model

The QCD model [3] defines a reference system in the HSL color space (a variant of HSV) for qualitative color description, which is built according to Figure 6 and defined as:

$$QC_{RS} = \{uH, uS, uL, QC_{NAME1..5}, QC_{INT1..5}\}$$

where uH is the unit of Hue; uS is the unit of Saturation; uL is the unit of Lightness; $QC_{NAME1..5}$ refers to the color names; and $QC_{INT1..5}$ refers to the intervals of HSL coordinates associated with each color. The chosen QC_{NAME} and QC_{INT} are:

$$QC_{NAME1} = \{\text{black}, \text{darkgrey}, \text{grey}, \text{lightgrey}, \text{white}\}$$

$$QC_{INT1} = \{[0, 20], [20, 30], [30, 50], [50, 75], [75, 100]\}$$

$$\in uL \mid \forall uH \wedge uS \in [0, 20]$$

$$QC_{NAME2} = \{\text{red}, \text{orange}, \text{yellow}, \text{green}, \text{turquoise}, \text{blue}, \text{purple}, \text{pink}\}$$

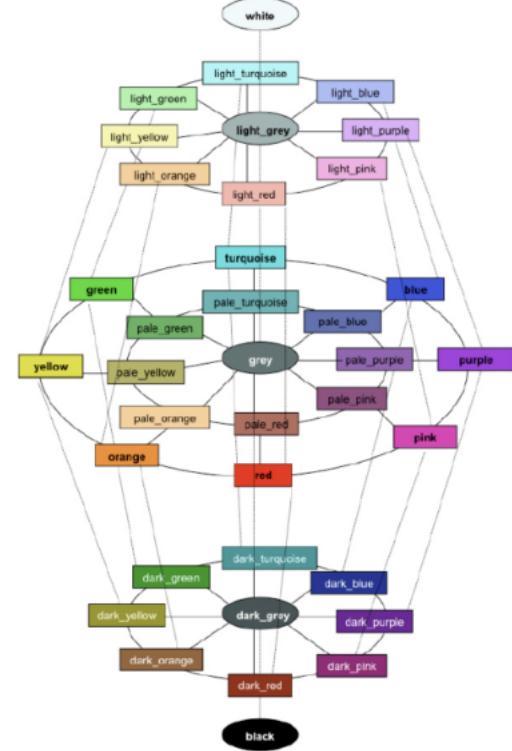


Figure 6. QCD color model

$$QC_{INT2} = \{(335, 360] \wedge [0, 20], [20, 50], [50, 80], [80, 160], [160, 200], [200, 260], [260, 300], [300, 335] \in uH \mid uS \in (50, 100] \wedge uL \in (40, 55]\}$$

$$QC_{NAME3} = \{\text{pale-red}, \text{pale-orange}, \text{pale-yellow}, \dots, \text{pale-blue}, \text{pale-purple}, \text{pale-pink}\}$$

$$QC_{INT3} = \{\forall QC_{INT2} \mid uS \in (20, 50] \wedge uL \in (40, 55]\}$$

$$QC_{NAME4} = \{\text{light-red}, \text{light-orange}, \text{light-yellow}, \dots, \text{light-blue}, \text{light-purple}, \text{light-pink}\}$$

$$QC_{INT4} = \{\forall QC_{INT2} \mid uS \in (50, 100] \wedge uL \in (55, 100]\}$$

$$QC_{NAME5} = \{\text{dark-red}, \text{dark-orange}, \text{dark-yellow}, \dots, \text{dark-blue}, \text{dark-purple}, \text{dark-pink}\}$$

$$QC_{INT5} = \{\forall QC_{INT2} \mid uS \in (50, 100] \wedge uL \in (20, 40]\}$$

In summary, the QCD defines two set of basic color labels (one monochromatic, one chromatic), which can be combined with “adjectives” (dark, light, pale) that capture meaningful variations in saturation and lightness in an intuitive way.

3.2 QCD interpretation of the KAN model

We can use the QCD to provide a qualitative interpretation of the transformation learned by the KANs. For each pixel in the original image which is covered by the watermarking mask, we compute the corresponding color as transformed by the learned KAN model. Then, we find the QCD label of both colors, thus obtaining a qualitative version of the mapping from the original to the watermarked colors. The results are shown in Table 1. The label on left side of the arrow represents the color in the original image, and the label on the right side represents the corresponding color on the watermarked image. Note that the mapping is not always one to one; in some cases,

some qualitative colors on the original image map to different qualitative labels in the watermarked image.

Table 1. Mapping of QCD colors under the watermarking transformation

<i>black</i>	\mapsto	<i>grey</i>
<i>dark_green</i>	\mapsto	<i>light_grey</i>
<i>dark_grey</i>	\mapsto	<i>grey light_grey</i>
<i>dark_yellow</i>	\mapsto	<i>light_grey</i>
<i>grey</i>	\mapsto	<i>light_grey</i>
<i>light_blue</i>	\mapsto	<i>white</i>
<i>light_green</i>	\mapsto	<i>white</i>
<i>light_grey</i>	\mapsto	<i>white light_grey</i>
<i>light_red</i>	\mapsto	<i>light_red white light_grey</i>
<i>light_yellow</i>	\mapsto	<i>white light_grey</i>
<i>pale_blue</i>	\mapsto	<i>light_grey</i>
<i>pale_green</i>	\mapsto	<i>light_grey</i>
<i>pale_red</i>	\mapsto	<i>light_grey</i>
<i>pale_yellow</i>	\mapsto	<i>light_grey</i>
<i>white</i>	\mapsto	<i>white</i>

Note how, in this case, some patterns are obvious:

- First of all, by using a qualitative representation the color labels are immediately understandable. For example, RGB coordinates (1, 0.14, 0.19) or HSV coordinates (357 deg, 0.85, 1), and all perceptually similar sections of the color space are just referred to by the natural language label “red”.
- By examining the table, we can see that the transformation corresponds to a “lightening” of the colors, transforming *dark* colors to their *light* versions, and converting *light* colors to *white* in some cases. Thus, the interpretation becomes immediately obvious.
- Also note that some colors are transformed to labels in the gray scale. This corresponds to a well-known feature of the human vision system, modeled by the QCD, in which very muted colors are perceived as gray (i.e. their chromatic hue is lost), even though if we examine the quantitative coordinates of such colors the hue is unchanged.
- Finally, note how these descriptions correspond to the way in which a human would describe the difference between the watermarked image and the original one. The watermarked areas are normally thought of as “whitened”, “lightened” or “muted” with respect to the original version.

These reasons illustrate why qualitative representations are an excellent fit for model explainability in general. And, in this particular case, they provide a natural complement to the symbolic expressions learned by the KAN.

4 Discussion: enhancing the interpretability of KANs with qualitative techniques

In the previous section we have shown how a qualitative representation can be used to provide an intuitive explanation of the result of a model. This is called a post-hoc explanation, and it can certainly be applied to models other than a KAN. However, KANs have some specific properties that make them especially interesting to be used in combination with qualitative techniques. Here we provide two promising examples:

First of all, remember that KANs depend on a suitable family of basis functions to be fitted. In the base implementation we have used in this paper, these functions are B-splines. However, in principle many other function basis are possible; some that have been recently tried are e.g. radial basis functions [6] and wavelets [1]. While using

qualitative functions directly is not possible since they are not differentiable, it is indeed possible to use some basis functions that are more readily interpretable in a qualitative way, such as fuzzy basis functions [4].

Another aspect in which qualitative approaches are potentially useful is as *constraints*. It’s possible to incorporate constraints to guide the training of the a KAN; this has been used, for example, to incorporate physical knowledge into the system. Of course, there is a long tradition in the field of qualitative reasoning of defining qualitative theories to be used in this way, and in principle it should be possible to incorporate domain knowledge into KAN training using qualitative reasoning techniques.

We consider that these aspects are promising avenues of research that combine the strengths of KAN and qualitative reasoning techniques.

5 Conclusion

In this paper we have introduced Kolmogorov-Arnold networks, and how they can be used to obtain symbolic approximations of functions. After applying them to a simple case study, we have seen how these symbolic formulas can be hard to interpret. As a solution, we have applied the QCD qualitative color theory to find an intuitive explanation of the result. Finally, we have introduced several topics for further research: finding basis functions which are suitable for generating qualitative interpretations, and the incorporation of qualitative constraints into the training process. We consider that KANs and qualitative approaches are complementary approaches, and that these directions of research may provide useful results.

Acknowledgements

This research has been partially funded by the Spanish Ministry of Science under grants PID2021-123152OB-C22 and PDC2021-121097-I00 both funded by the MCIN/AEI/10.13039/501100011033.

References

- [1] Z. Bozorgasl and H. Chen. Wav-KAN: Wavelet Kolmogorov-Arnold networks, 2024.
- [2] M. Cheon. Kolmogorov-arnold network for satellite image classification in remote sensing, 2024.
- [3] Z. Falomir, L. Museros, and L. Gonzalez-Abril. A model for colour naming and comparing based on conceptual neighbourhood. an application for comparing art compositions. *Knowledge-Based Systems*, 81:1–21, 2015.
- [4] H. M. Kim and J. Mendel. Fuzzy basis functions: comparisons with other basis functions. *IEEE Transactions on Fuzzy Systems*, 3(2):158–168, 1995. doi: 10.1109/91.388171.
- [5] A. Kolmogorov. On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables. *Proceedings of the USSR Academy of Sciences*, pages 179–182, 1956. English translation: Amer. Math. Soc. Transl., 17 (1961), pp. 369–373.
- [6] Z. Li. Kolmogorov-Arnold networks are radial basis function networks, 2024.
- [7] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark. KAN: Kolmogorov-Arnold Networks, 2024.
- [8] L. Longo, M. Brčic, F. Cabitzza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, and S. Stumpf. Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024.
- [9] C. J. Vaca-Rubio, L. Blanco, R. Pereira, and M. Caus. Kolmogorov-arnold networks (kans) for time series analysis, 2024.