Evaluating the Interpretability of the Knowledge Compilation Map: Communicating Logical Statements Effectively

Serena Booth[1], **Christian Muise**[2,3] and **Julie Shah**[1]
[1]MIT Computer Science and Artificial Intelligence Laboratory
[2]IBM Research
[3]MIT-IBM Watson AI Lab
{serenabooth, julie_a_shah}@csail.mit.edu, christian.muise@ibm.com

Abstract

Knowledge compilation techniques translate propositional theories into equivalent forms to increase their computational tractability. But, how should we best present these propositional theories to a human? We analyze the standard taxonomy of propositional theories for relative *interpretability* across three model domains: highway driving, emergency triage, and the chopsticks game. We generate decision-making agents which produce logical explanations for their actions and apply knowledge compilation to these explanations. Then, we evaluate how quickly, accurately, and confidently users comprehend the generated explanations. We find that domain, formula size, and negated logical connectives significantly affect comprehension while formula properties typically associated with interpretability are not strong predictors of human ability to comprehend the theory.

1 Introduction

Given a propositional theory and a set of queries to resolve, *knowledge compilation* techniques translate the theory to a target compilation representation. This translation is typically expensive, but, if the compiled form is well suited to its application, subsequent querying and transformations can be guaranteed to be efficient. Knowledge compilation is useful in AI system diagnosis and state estimation.

Darwiche and Marquis introduced the knowledge compilation map, which relates each logical form to its succinctness and tractable computations [2002]. They argued that some logical languages are *representation* languages suitable for humans to read and write, while other logical languages are *target compilation* languages. They claimed that neither the intersection nor the union of these sets is null; however, they did not formally define representation languages. To our knowledge, there has been no systematic study of the separation of representation and compilation languages. Alongside the succinctness and the class of queries and transformations the language can support in polytime, we propose extending the knowledge compilation map to consider the relative *interpretability* of each form.

While it is typically unreasonable to present a large logical formula to a human user, the question of how to best present even a small formula remains open. There are two principle unresolved questions in how knowledge compilation relates to human cognition: is it worth the expense of translating a formula to an alternate representation before presenting it to a human user? And, are the same properties which enable tractable machine computation also useful for human computation? We conduct a user study to evaluate whether knowledge compilation can aid logic interpretability. We find only sparse effects of knowledge compilation properties on interpretability. We discover some languages considered to be compilation-only are acceptable, while disjunctive normal form—a representation assumed to be interpretable—is not significantly more interpretable than other forms.

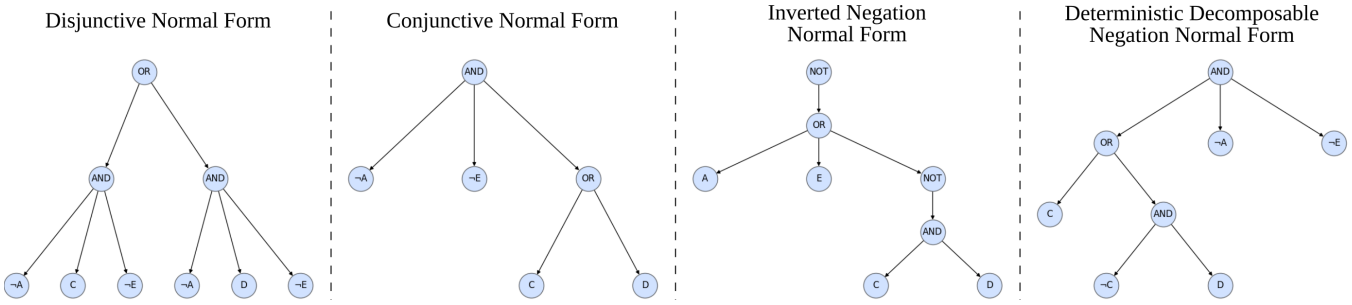


Figure 1: We evaluate the interpretability of equivalent propositional theories, here shown as directed acyclic graphs where each connective and each variable is a node. NOT can be represented as a node with an edge to the connective or to the variable it is negating, or it can be represented through the connectives NAND, NOR, or directly applied to a variable (e.g. $\neg A$). In disjunctive normal form (left), this statement is written $(\neg A \wedge C \wedge \neg E) \vee (\neg A \wedge D \wedge \neg E)$, where A,C,D, and E are variables.

2 Background

2.1 Explainable Systems

The importance of explainability is exemplified by GDPR, wherein the European Union ruled that users have a "right to an explanation" [Goodman and Flaxman, 2017]. However, there is currently no consensus on precise definitions of *interpretability* or *explainability* [Lipton, 2016; Gilpin *et al.*, 2018]. Doshi-Velez and Kim presented a taxonomy of interpretability evaluation which supports using "human-grounded metrics" to evaluate the general quality of an explanation [2017]. We adhere to their taxonomy, and, in line with prior studies [Huysmans *et al.*, 2011], we define the relevant metrics for interpretability of propositional theories to be accuracy, speed, and confidence of interpretation. We claim that individual propositional theories can be evaluated for their interpretability, while a system which reasons over propositional theories is explainable if every reachable state and action can be described through such an interpretable statement.

Absent a precise definition of interpretability, succinctness could be considered as a proxy metric. Though the amount of information that humans can process is constrained, i.e., 7 ± 2 cognitive entities at once [Miller, 1956], below this the succinctness proxy metric is less clear. Further, the literature demonstrates that this view of interpretability is misleading: Freitas argued against using model size as the only criteria for interpretability, and discussed the usefulness of monotonicity constraints in which the value of a predictor attribute monotonically increases or decreases the probability of class membership [2014].

2.2 Propositional Theories as Explanations

Knowledge compilation has proven useful in domains where explainable AI collaboration with a human is essential. One compiled representation has been used for system state estimation [Barrett, 2005; Elliott and Williams, 2006], and the task of system diagnosis has been accomplished by making use of several languages from the knowledge compilation map [Sztipanovits and Misra, 1996; Torasso and Torta, 2006; Huang and Darwiche, 2005; Siddiqi and Huang, 2011].

Motivated by the conventional wisdom that disjunctive normal form (DNF) is an interpretable representation, Hayes and Shah demonstrated a system which summarizes embedded control logic as a DNF explanation [2017]. An example explanation from a robot part inspection task is: "I look in the middle when the stock feed is off or when the stock feed is on and I have not detected a part and I am looking high." Wang et al. likewise built DNF models for classification tasks under the premise that such a model should be interpretable to experts [2015]. Due in part to its expressiveness, DNF has also become standard for decision models in the marketing community [Hauser *et al.*, 2010].

Miller argued that any explanations should be contrastive and include only the most pertinent information [2018]. Propositional theories are well suited to these properties and offer expressive flexibility. However, prior works raise the questions of whether DNF is always the most expressive and interpretable representation, and whether an alternate knowledge compilation form enables both scalability and interpretability. To help answer these questions, this work aims to extend the knowledge compilation map to understand the interpretability of each form.

2.3 Presentation Formats and Interpretability

Huysmans et al. compared the interpretability of decision tables, binary decision trees, and propositional rules for the domain of evaluating credit applications [2011]. Using the metrics of accuracy, speed, and confidence of interpretation as a proxy for interpretability, they found that decision tables were the most interpretable for their 51 non-expert participants. Subramanian et al. [1992] and Allahyari and Lavesson [2011] similarly compared the interpretability of decision tree models and rule-based models and determined that decision trees were more interpretable. Several factors may contribute to this discrepancy. Each study defined interpretability differently: Huysmans et al. combined accuracy, speed, and confidence; Allahyari and Lavesson used perceived understandability; and Subramanian et al. used accuracy only. These studies' populations also differed: Huysmans et al. considered non-experts while Subramanian et al. and Allahyari and Lavesson considered experts.

Our work presents propositional theories as text sentences. While interpretability could potentially be increased by exploring alternative presentations, we use text to establish a baseline of interpretability across classes of propositional theories.

3 Knowledge Compilation

3.1 Properties

Any propositional theory can be represented as a directed acyclic graph (see Figure 1). If Σ is a propositional theory, we define $\texttt{Vars}(\Sigma)$ to be all variables in Σ. Where C is any node in a graph, we define $\texttt{Vars}(C)$, to be the set of all variables used in composing C and its descendants. We define the properties:

- **Flat**. A sentence is flat if in its graph representation the distance from the root to any leaf is at most 2.
  - **Simple Disjunction**. A flat sentence has simple disjunction if it consists of a conjunction of disjunctive clauses. For example: $(A \vee B) \wedge (\neg A \vee D)$.
  - **Simple Conjunction**. A flat sentence has simple conjunction if it consists of a disjunction of conjunctive clauses. For example: $(A \wedge B) \vee (\neg A \wedge D)$.
- **Decomposable**. A sentence is decomposable if, for each AND node in the graph, its children do not share variables. If AND node C has children $c_0, c_1, ..., c_n$, then $\forall i, j \in [0, n]$, where $i \neq j$, $\texttt{Vars}(c_i) \cap \texttt{Vars}(c_j) = \emptyset$.
- **Deterministic**. A sentence is deterministic if, for each OR node in the graph, its children are logically contradictory. If OR node C has children $c_0, c_1, ..., c_n$, then $\forall i, j \in [0, n]$, where $i \neq j$, $c_i \wedge c_j \equiv \texttt{False}$.
- **Decisive**. A sentence is decisive if each OR node in the graph is a decision node.
  - **Decision Node**. A node labeled True or False, or an OR node with the form $(X \wedge \alpha) \vee (\neg X \wedge \beta)$, where X is a *decision variable* and α, β are nodes. If α or β is an OR node, it must be a decision node. α or β may be NULL or AND nodes.
- **Ordered**. A sentence is ordered if, on every path from the root to each leaf, decision variables follow the same order.

- **Smooth**. A sentence is smooth if, for each OR node in the graph, its children are composed of the same set of variables. If OR node C has children $c_0, c_1, ..., c_n$, then $\forall i,j \in [0,n], \texttt{Vars}(c_i) = \texttt{Vars}(c_j)$.

The above properties only apply to sentences in which negation is *not* applied to logical connectives. As this may be a limitation of the interpretability of these theories, we define:

- **ε-inverted**. A sentence is ε-inverted if children of OR and AND nodes have no more than an ε ratio of negated nodes.
- **Inverted**. A sentence is inverted if it meets the ε-inverted criteria for $\varepsilon = 1/2$. The sentence "$A \wedge \neg(B \vee C)$" is inverted, but the sentence "$A \wedge \neg B \wedge \neg C$" is not.

3.2 Languages

Based on these properties, we define the logical languages:

- **NNF: Negation Normal Form.** A sentence in which negation is only applied to literals and not to logical connective (AND or OR) nodes.
- **DNF: Disjunctive Normal Form.** Every NNF sentence which is flat and satisfies the criteria for simple conjunction.
- **CNF: Conjunctive Normal Form.** Every NNF sentence which is flat and satisfies the criteria for simple disjunction.
- **ODNF: Orthogonal Disjunctive Normal Form.** Every NNF sentence which is deterministic, flat, and satisfies the criteria for simple conjunction. Deterministic DNF.
- **d-DNNF: Deterministic Decomposable Negation Normal Form.** Every NNF sentence which is deterministic and decomposable.
- **sd-DNNF: Smooth Deterministic Decomposable Negation Normal Form.** Every NNF sentence which is smooth, deterministic, and decomposable.
- **OBDD: Ordered Binary Decision Diagram.** Every NNF sentence which is ordered, decisive, and decomposable.
- **MODS: Models**. Every NNF sentence which is deterministic, smooth, and satisfies the criteria for simple conjunction.
- **INNF: Inverted Negation Normal Form.** Every sentence which meets the inverted criteria. Note this is not a subset of NNF.

While this list describes the set of languages we set out to evaluate, we ultimately removed two languages (and their corresponding properties) due to incomprehensibly large formulae: during our pilot studies, participants were unwilling to engage with the lengthy sd-DNNF and MODS representations.

4 Testing Interpretability of Logical Sentences

We conducted a user study to evaluate the interpretability of logical sentences of different forms. We evaluate human response to AI agents in three model domains: (1) highway driving, (2) emergency triage based on one of Hodgetts and Porter's sieve procedures [2002], and (3) chopsticks, a combinatorial hand game similar in nature to Tic-Tac-Toe (Figure 2). For each domain, we create a "good" agent which behaves in an intuitive but not necessarily optimal manner, and a "bad" agent which behaves in a faulty or unintuitive manner.

4.1 Generating and Translating Explanations

Using the technique of Hayes and Shah [2017], we define predicates of interest as annotations to the control logics of these AI agents. For example, in the highway driving scenario, the natural language form of the predicates may be {vehicle to my left, vehicle to my right, vehicle in front of me, vehicle behind me, my exit is next}. The control logic of the agent may be rule-based or sub-symbolic. We simulate the agents and generate traces detailing each state experienced, action undertaken, and new state encountered by the agent: (s,a,s'). Given a state, we can look up the truth assignment to the annotated predicates. From these traces, we generate disjunctive normal form control logic summaries for each behavior.

For the highway driving scenario in which the agent has the action set {slow down, speed up, merge left, merge right, do nothing}, we might query "When do you merge left?" In response, the good agent may provide the summary "when not a vehicle is to my left and a vehicle is in front of me —or— not a vehicle is to my left and a vehicle is to my right." The bad agent may respond, "when there is a vehicle to my left and my exit is not next —or— not a vehicle is in front of me." In this scenario, the bad agent could cause a crash on the road.

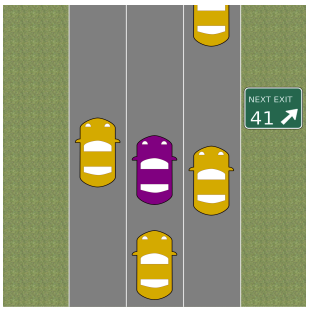
For each action-agent pair, we query "when do you a" and create DNF explanations. We convert each DNF representation into CNF by applying the distributive laws, then we simplify the resulting CNF representation with the PMC PREPROCESSOR [Lagniez and Marquis, 2014]. We convert these CNF representations into d-DNNF using DSHARP [Muise *et al.*, 2012]. We directly convert to ODNF, OBDD, and INNF forms. In total, we consider DNF, CNF, d-DNNF, INNF (converted from DNF), INNF (converted from CNF), INNF (converted from d-DNNF), ODNF, and OBDD.

For each "good" agent and each representation, we present two scenarios: one with an action the agent would take and one with an action the agent would not take. For each "bad" agent, we select the subset of languages {INNF (converted from CNF), d-DNNF, DNF, and ODNF}, and randomly selected actions which the agent either would or would not take. Our interleaved "bad" agents prevent users from relying on domain intuition to resolve logical formulae and instead require comprehension of the presented logical explanations.

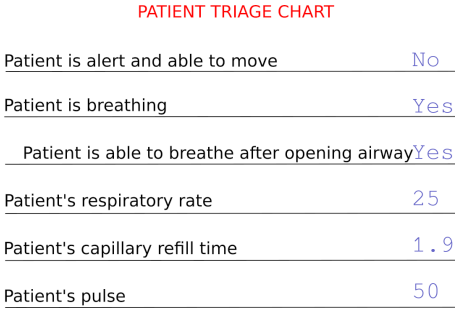
4.2 User Study Presentation

We present participants with natural language forms of these explanations. We replace logical connective "and" with "both" or "all of" depending on the number of child nodes; "or" with "one or both of" or "one or more of"; "nand" with "not both of" or "not all of"; and "nor" with "neither of" or "none of." We also push negation into the natural position in the predicate, e.g., "a vehicle is not to my left" instead of "not a vehicle is to my left." We present the formulae as bulleted indented lists (Figure 2).

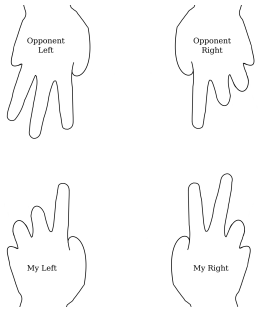
Over the duration of the study, participants answer 60 scenario questions. Domains were presented in random order, with questions randomized per domain. At the beginning of each domain section, participants read instructions and responded to a sample question. For each scenario and formula presented, we record accuracy, time spent, and a 5-point Likert scale measure of confidence in their answer. To measure overall workload, at the end of each domain section we present the participant with



I speed up when
- One or both of:
  - Both of:
    - a vehicle is not in front of me
    - the next exit is not 42
  - All of:
    - a vehicle is to my right
    - a vehicle is not in front of me
    - a vehicle is behind me



I triage as urgent when
- All of:
  - patient is breathing
  - respiratory rate is between 10 and 30
  - One or both of:
    - pulse is between 40 and 70
    - Both of:
      - pulse is not between 40 and 70
      - capillary refill takes less than 2 seconds



I add my right hand to my opponent's left hand when
- All of:
  - One or both of:
    - my left hand is out
    - my right hand outs my opponent's left
  - One or both of:
    - my right hand is not out
    - my right hand outs my opponent's left
  - One or both of:
    - my right hand does not out my opponent's left
    - my right hand outs my opponent's left

Figure 2: Example scenarios from each domain with behavior explanations. Participants must decide whether (1) the purple car will speed up, (2) to triage as urgent, and (3) to add their right hand to their opponent's left hand. In these example explanations, the car scenario is in DNF, the triage scenario is in d-DNNF, the hand game is in CNF, and all propositional theories resolve to true.

a NASA raw-TLX questionnaire [Hart and Staveland, 1988; Grier, 2015], with a scale set from 0 (very low) to 100 (very high). Finally, participants completed a qualitative questionnaire and were invited to discuss their experiences interpreting the logical forms.

4.3 Hypotheses

As discussed, DNF is considered the de facto representation for building interpretable models. Due to the popularity and multi-discipline convergence on this representation, we hypothesize:

- **[H1]** Participants will resolve the truth value of DNF representations more accurately, more confidently, and faster.

Propositional theory predicates can be abstracted to variables, and between domains these underlying representations can be identical. However, the effort required to parse a natural language predicate may vary by domain. For example, in the emergency triage domain, participants must evaluate mathematical inequalities to resolve the predicates. Thus, we hypothesize:

- **[H2]**: Domain and Interpretability
  - **[H2_accuracy]** Domain *will not* affect accuracy in resolving the truth value of a propositional theory.
  - **[H2_confidence]** Domain *will not* affect confidence in resolving the truth value of a propositional theory.
  - **[H2_time]** Domain *will* affect time spent resolving the truth value of a propositional theory.

Lastly, we consider the existence of the contraction "nor" in natural language and in contrast the lack of the word "nand" [van Wijk, 2006]. Due to familiarity with "nor" and lack of familiarity with "nand," we hypothesize:

- **[H3]** The presence of "nor" will not affect accuracy, confidence, or speed in resolving propositional theories. The presence of "nand" will decrease interpretability.

5 Analysis

5.1 Demographics

We recruited 25 participants from a local university, MIT. Participants completed the study on-site in approximately one hour. The study was within-subjects; all participants considered all domains and all questions. Age ranged from 19 to 66 ($\mu = 28, SD = 10$). 17 participants identified as female; 8 as male. Participants ranged in education level from high school graduates to PhDs; most commonly, participants had Bachelor's degrees $(11/25)$. On a 5-point scale to evaluate familiarity with logic, participants provided a self-assessed mean of 3.6 (between neutral and some experience), with standard deviation 0.9. Most participants $(18/25)$ identified as native English speakers. The study size and population must be kept in mind when considering the results presented here.

5.2 Measures & Methods

- **Accuracy.** Did participants correctly evaluate the truth value of a logical statement?
- **Self-reported confidence.** "How confident are you of your answer?" Participants responded on a 5-point Likert scale, from "1—very unconfident" to "5—very confident."
- **Time.** How long participants spent evaluating each question measured as time to webpage submit.

We code accuracy as a binary variable, confidence as a categorical variable, and time as a continuous variable. As the relative weightings of these proxy measures of interpretability are unknown, we construct a model for each measure. For our analysis, we use generalized linear mixed models which allow us to account for our repeated measures design [Kaptein, 2016]. The expectation is that any given participant will be consistently faster or slower, more or less accurate, and more or less confident over the course of the evaluation when compared with the larger population. ANOVA is only applicable when the measured variable is continuous; as such, it cannot be used

Fixed Factor	Accuracy			Time			Confidence		
	χ^2	df	p	χ^2	df	p	χ^2	df	p
Domain	21.118	2	<0.001	74.533	2	<0.001	124.08	2	<0.001
Formula Size	0.005	8	0.946	66.047	8	<0.001	15.767	8	<0.001
Inverted	3.975	1	0.046	0.061	1	0.806	0.390	1	0.532
Simple Disjunction	0.043	1	0.836	0.154	1	0.695	0.638	1	0.424
Simple Conjunction	0.284	1	0.594	0.111	1	0.739	0.518	1	0.472
Deterministic	0.451	1	0.502	0.151	1	0.698	0.427	1	0.513
Decomposable	0.000	1	0.996	1.830	1	0.176	4.239	1	0.040
Decisive	0.884	1	0.347	5.015	1	0.025	0.070	1	0.404

Table 1: Generalized linear mixed model comparisons. Domain affects accuracy, time, and confidence, while formula size affects time and confidence. Though some language properties show significant effects, such as invertedness on accuracy, most properties do not.

to predict accuracy or confidence. We use repeated measures ANOVA to analyze the effect of each domain on the perceived workload, measured through the NASA-RTLX.

For the generalized linear mixed models to predict accuracy, confidence, and time, each language property (ε-inverted, simple disjunction, simple conjunction, deterministic, decomposable, and decisive) is considered to be a fixed effect. This set is restricted to non-overlapping properties for our selected languages; flat, smooth, and ordered are removed. In addition, some generated explanations may be larger than others, and so we consider the size of the formula as a fixed effect (measured as the number of nodes in the directed acyclic graph representation). Lastly, the domain is considered as a fixed effect. The participant identifier is a random effect, allowing us to account for the different subject baselines. To evaluate each fixed effect, we perform likelihood ratio tests of two model variants: one including the fixed effect of interest and one not. In all analysis, we report results for $\alpha = 0.05$. Table 1 summarizes these models.

Interpretability Proxy: Accuracy

Overall, participants achieved an average accuracy rate of 92.6% ($SD = 9.1\%$). In our generalized linear mixed models for accuracy, we find the significant fixed factors to be domain ($\chi^2(2, 1492) = 21.118, p < 0.001$) and inversion ($\chi^2(1,1492) = 3.975, p = 0.046$). Notably, the presence of simple conjunction, the fixed effect which distinguishes Disjunctive Normal Form, was not a significant predictor of accuracy ($\chi^2(1,1492) = 0.284, p = 0.594$). In constrast to our hypothesis **[H1]**, this suggests that languages other than DNF may be as interpretable from the perspective of accuracy. The significance of inversion relates to our hypothesis **[H3]**, though inverted formulae here include both nor and nand logical connectives.

Interpretability Proxy: Time

On average, participants answered each question in 32.5 seconds ($SD = 27.8$). Figure 3 shows the impact of each fixed effect on time. We find the significant fixed factors in our generalized linear mixed model to be domain ($\chi^2(2,1492) = 74.533, p < 0.001$), formula size ($\chi^2(8,1492) = 66.047, p < 0.001$), and decisiveness ($\chi^2(1,1492) = 5.015, p = 0.025$). In contrast to our hypothesis **[H1]**, we find simple conjunction is not a significant predictor for time spent resolving propositional theories ($\chi^2(1,1492) = 0.111, p = 0.739$).

Interpretability Proxy: Confidence

On average, participants were "very confident" across all questions ($M = 4.56, SD = 0.84$). From our generalized linear mixed models, we find the significant predictors of confidence to be domain ($\chi^2(2, 1492) = 124.08, p < 0.001$), formula size ($\chi^2(8,1492) = 15.767, p < 0.001$), and decomposability ($\chi^2(1,1492) = 4.239, p = 0.040$). Once more, we find evidence against **[H1]**, as simple conjunction is not a significant predictor for confidence ($\chi^2(1,1492) = 0.518, p = 0.472$).

Adding accuracy as a fixed effect in our generalized linear mixed model for confidence shows accuracy to be a strong predictor ($\chi^2(1,1492) = 59.942, p < 0.001$). Similarly, if we add time as a fixed effect, it too is a strong predictor of confidence ($\chi^2(1, 1492) = 10.186, p = 0.001$). Intuitively, when participants are less confident, they take longer and are less accurate when resolving the truth value of a propositional theory.

Impact of Different Domains

Across domains, task complexity varies. The highway domain requires a spatial frame of reference (left, right, front, back). The emergency domain requires evaluation of inequalities (e.g., "pulse is not between 40 and 70"). The chopsticks domain requires a spatial frame of reference (left, right) and addition. We compare RTLX-measured workload across the domains using repeated measures ANOVA. We find we likely can reject the null hypothesis that the domain did not affect the experienced workload ($F(1.656, 368.574) = 6.882, p = 0.004$). In pairwise comparisons, we see a significant difference between the chopsticks domain ($M = 40.35$, $SD = 12.23$) and the highway domain ($M = 33.74$, $SD = 12.20$; $p = 0.016$), and between the chopsticks and emergency triage domains ($M = 35.09$, $SD = 11.35$; $p = 0.046$). We do not see a significant difference between the highway and emergency triage domains.

We further consider domain, modeled as a fixed effect, in the generalized linear mixed models for accuracy, time, and confidence. We find domain significantly affects all three metrics: accuracy ($\chi^2(2,1492) = 21.118, p < 0.001$), time ($\chi^2(2,1492) = 74.533, p < 0.001$), and confidence ($\chi^2(2,1492) = 124.08, p < 0.001$). While the correlation between higher experienced workload and increased time spent resolving propositional theories is expected and supports **[H2_time]**, accuracy and confidence likewise drop in contrast with **[H2_accuracy], [H2_confidence]**. As these metrics collectively proxy interpretability, a change in one metric may affect the other metrics.

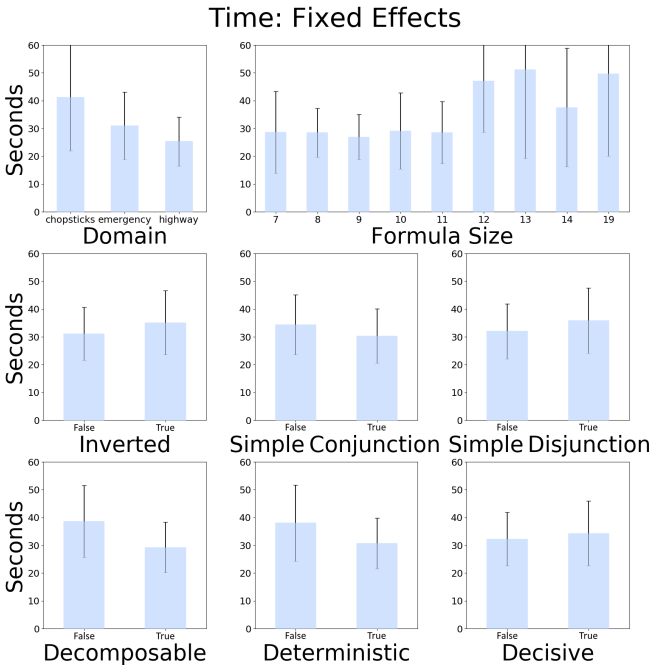


Figure 3: Time spent per domain, per formula size, and per language property, normalized by individual participant performance. Only domain, formula size, and decisiveness have significant effects on time. Error bars are standard deviation.

5.3 Qualitative Results & Discussion

Disjunctive Normal Form

Conventional wisdom presumes an interpretable propositional theory should be written either in Disjunctive Normal Form [Hayes and Shah, 2017; Wang *et al.*, 2015; Hauser *et al.*, 2010] or, perhaps, in Conjunctive Normal Form [Darwiche and Marquis, 2002]. Our analysis suggests there is more flexibility in the presentation of propositional theories than previously assumed. Neither simple disjunction nor simple conjunction is a significant fixed effect in predicting the accuracy, time, or confidence in resolving a propositional theory (Table 1). While interpretability varied across domains and inversion, it appears somewhat robust across the tested subset of the knowledge compilation map. Decomposability resulted in a statistically significant increase in confidence, while decisiveness resulted in a statistically significant increase in time spent.

In follow-up discussions with participants, we showed four language examples (DNF, INNF, CNF, and d-DNNF). Most participants identified either the DNF or the CNF form as the "easiest." Nonetheless, when specifically asked about the d-DNNF form with four layers of nesting (e.g., the longest path from the root of the directed acyclic graph to a leaf was 4), participants said it would be "not too hard; but it may take longer due to nesting," or that it may be "tedious but easy to follow." Participants discussed the trade-offs between the different forms; for example, many participants mentioned that the DNF form allowed them to stop evaluating the theory earlier, while the d-DNNF forms sometimes required even less evaluation.

The Importance of Domain

In follow-up discussions, $17/25$ participants identified the chopsticks hand game as the hardest domain. Their explanations ranged from "it came last so I was tired" to "it came first, and was challenging to understand the framework." Some participants mentioned that having to perform arithmetic increased the difficulty of the task; other participants mentioned the difficulty in keeping track of the frame of reference; others still mentioned the complexity of the rule set. These comments are supported by the results of the evaluation of the NASA-RTLX workload scores, as well as by all three metrics of interpretability. This suggests that differences in domains may require different presentation of propositional theories, including the determination of potential costs associated with resolving predicates.

Negation Applied to Logical Connectives

In follow-up discussions with participants, we asked, "Which, if any, logical connectives were hard?" $23/25$ participants said that "none of" was particularly challenging, while only one participant explicitly mentioned "not both of." Due to the frequency with which participants responded that "none of" was the hardest logical connective, study operators stopped listing it. In spite of this, participants continued to identify it as the most challenging connective, contrasting with **[H3]**. Participants described "none of" as "counterintuitive," and explained that it required them to "flip the logic." One participant explained that, as they are "usually looking for true statements, none of was especially hard to track." $10/25$ participants also identified double-negatives as challenging. When viewing an INNF example with two negated logical connectives, one participant exclaimed, "Yikes, I hate negatives!" In cognitive psychology, Chase and Clark found that participants were slower to resolve the truth value of a single predicate when it was negated [1972]. Both our qualitative and quantitative results suggest explanations should aim to avoid inverted framing wherever possible.

Knowledge Compilation and Human Cognition

Darwiche and Marquis map the *queries* and *transformations* each logical language supports in polytime (assuming P $\neq$ NP) [2002]. Ordered by query expressivity, CNF supports the fewest polytime queries while OBDD supports the most: CNF $<$ ODNF $\leq$ DNF $<$ d-DNNF $\leq$ OBDD. INNF is not evaluated. Considering human cognition, decomposability (of d-DNNF, sd-DNNF, and OBDD) had a significant effect on increased confidence, while decisiveness (of OBDD) had a significant effect on increased time spent. Increased computational tractability appears to neither help nor hinder human computation.

6 Conclusion

We revisit the knowledge compilation map to investigate how to best present propositional theories to humans. We find properties associated with interpretability such as simple conjunction did not have significant effects, suggesting translations to these forms may be unnecessary. Further, we find sparse effects across other knowledge compilation properties; only inversion (on accuracy), decisiveness (on time spent), and decomposability (on confidence) had significant effects. Our work moves toward determining how logic can be used in the design of an interpretable language of AI. Our study procedures and source code are available at github.com/serenabooth/logic-interpretability.

References

[Allahyari and Lavesson, 2011] Hiva Allahyari and Niklas Lavesson. User-oriented assessment of classification model understandability. In *11th scandinavian conference on Artificial intelligence*. IOS Press, 2011.

[Barrett, 2005] Anthony Barrett. Model compilation for real-time planning and diagnosis with feedback. In *IJCAI*, 2005.

[Chase and Clark, 1972] William G Chase and Herbert H Clark. Mental operations in the comparison of sentences and pictures. *Cognition in learning and memory*, 1972.

[Darwiche and Marquis, 2002] Adnan Darwiche and Pierre Marquis. A knowledge compilation map. *Journal of Artificial Intelligence Research*, 17:229–264, 2002.

[Doshi-Velez and Kim, 2017] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[Elliott and Williams, 2006] Paul Elliott and Brian Williams. DNNF-based belief state estimation. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 36. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[Freitas, 2014] Alex A Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10, 2014.

[Gilpin *et al.*, 2018] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89. IEEE, 2018.

[Goodman and Flaxman, 2017] Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.

[Grier, 2015] Rebecca A Grier. How high is high? a meta-analysis of NASA-TLX global workload scores. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 59, pages 1727–1731. SAGE Publications Sage CA: Los Angeles, CA, 2015.

[Hart and Staveland, 1988] Sandra G Hart and Lowell E Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[Hauser *et al.*, 2010] John R Hauser, Olivier Toubia, Theodoros Evgeniou, Rene Befurt, and Daria Dzyabura. Disjunctions of conjunctions, cognitive simplicity, and consideration sets. *Journal of Marketing Research*, 47(3):485–496, 2010.

[Hayes and Shah, 2017] Bradley Hayes and Julie A Shah. Improving robot controller transparency through autonomous policy explanation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pages 303–312. ACM, 2017.

[Hodgetts and Porter, 2002] Timothy J. Hodgetts and Crispin Porter. *Major incident management system: the scene aid memoire for major incident medical management and support*. BMJ, 2002.

[Huang and Darwiche, 2005] Jinbo Huang and Adnan Darwiche. On compiling system models for faster and more scalable diagnosis. In *Proceedings Of The National Conference On Artificial Intelligence*, volume 20, page 300, 2005.

[Huysmans *et al.*, 2011] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, 2011.

[Kaptein, 2016] Maurits Kaptein. *Using Generalized Linear (Mixed) Models in HCI*, pages 251–274. 03 2016.

[Lagniez and Marquis, 2014] Jean-Marie Lagniez and Pierre Marquis. Preprocessing for propositional model counting. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[Lipton, 2016] Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[Miller, 1956] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[Miller, 2018] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018.

[Muise *et al.*, 2012] Christian Muise, Sheila A McIlraith, J Christopher Beck, and Eric I Hsu. D sharp: fast d-DNNF compilation with sharpSAT. In *Canadian Conference on Artificial Intelligence*, pages 356–361. Springer, 2012.

[Siddiqi and Huang, 2011] Sajjad Ahmed Siddiqi and Jinbo Huang. Sequential diagnosis by abstraction. *Journal of Artificial Intelligence Research*, 41:329–365, 2011.

[Subramanian *et al.*, 1992] Girish H. Subramanian, John Nosek, Sankaran P. Raghunathan, and Santosh S. Kanitkar. A comparison of the decision table and tree. *Commun. ACM*, 35(1):89–94, January 1992.

[Sztipanovits and Misra, 1996] Janos Sztipanovits and Amit Misra. Diagnosis of discrete event systems using ordered binary decision diagrams. In *Seventh International Workshop on Principles of Diagnosis*. Citeseer, 1996.

[Torasso and Torta, 2006] Pietro Torasso and Gianluca Torta. Model-based diagnosis through OBDD compilation: A complexity analysis. In *Reasoning, Action and Interaction in AI Theories and Systems*, pages 287–305. Springer, 2006.

[van Wijk, 2006] Maarten van Wijk. *Logical connectives in natural language: a cultural evolutionary approach*. PhD thesis, Universiteit Leiden, 2006.

[Wang *et al.*, 2015] Tong Wang, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry MacNeille. Or's of and's for interpretable classification, with application to context-aware recommender systems. *CoRR*, abs/1504.07614, 2015.