Abduction in Human Reasoning

Steffen Hölldobler and **Tobias Philipp** and **Christoph Wernhard**
International Center for Computational Logic
Technische Universität Dresden
01062 Dresden, Germany
sh@iccl.tu-dresden.de

Abstract

In this paper we contribute to bridging the gap between human reasoning as studied in Cognitive Science and commonsense reasoning based on formal logics and formal theories. In particular, the suppression task studied in Cognitive Science provides an interesting challenge problem for human reasoning based on logic. The work presented in the paper is founded on the recent approach by Stenning and van Lambalgen to model human reasoning with suppression by means of logic programs with a specific three-valued completion semantics and a semantic fixpoint operator that yields a least model, as well as abduction. Their approach has been subsequently made more precise and technically accurate by switching to three-valued Łukasiewicz logic. In this paper, we extend this refined approach by abduction. We show that the inclusion of abduction permits to adequately model additional empiric results reported from Cognitive Science. As a further extension, we discuss abduction with integrity constraints to model human reasoning. For the arising abductive reasoning tasks we give complexity results. Finally, we outline several open research issues that emerge from the application of logic to model human reasoning.

1 Introduction

In (McCarthy 1963) John McCarthy proposed a framework for reasoning about actions, causality, and causal laws, whose third postulate was that *the formal descriptions of situations should correspond as closely as possible to what people may reasonably be presumed to know about them when deciding what to do*. Human reasoning has been intensely studied within Cognitive Science (e.g. (Evans, Newstead, and Byrne 1993)) and there appears to be a widespread belief within the Cognitive Science community that logic is inadequate for human reasoning (e.g. (Byrne 1989)). Thus, an Artificial Intelligence approach to characterize commonsense reasoning using representations based on logic or other formal theories faces the formidable challenge of bridging the gap between human reasoning as studied within Cognitive Science and commonsense reasoning based on formal logics and formal theories.

Recently, in (Stenning and van Lambalgen 2008) Keith Stenning and Michiel van Lambalgen have proposed a two-stage process to model human reasoning. Given a sentence in natural language, the first step consists of reasoning towards an appropriate logical representation, whereas in the second step conclusions are drawn with respect to the models of the generated logical representations. They propose to use logic programs, strong Kleene three-valued semantics with strong equivalence (Kleene 1952), a certain variant of completion semantics, a semantic fixpoint operator which yields a least model as well as abduction. Furthermore, they demonstrate the adequateness of their proposal by showing how the various scenarios considered in Byrne's suppression task (Byrne 1989) are adequately modeled.

Unfortunately, the technical results of (Stenning and van Lambalgen 2008) contain an error, which was corrected in (Hölldobler and Ramli 2009b; 2009c) by considering the three-valued Łukasiewicz logic (Łukasiewicz 1920) instead of Kleene logic. However, the approach in (Hölldobler and Ramli 2009b; 2009c) does not include abduction and, consequently, some scenarios of Byrne's suppression task are not yet covered. In this paper we close this gap by adding abduction to the approach in (Hölldobler and Ramli 2009b; 2009c).

The paper is organized as follows: In Section 2 we will briefly present the suppression task as a challenge problem for human reasoning based on logic. In Section 3 we review the approach presented in (Stenning and van Lambalgen 2008) with the modifications discussed in (Hölldobler and Ramli 2009b; 2009c). We extend this approach by abduction in Section 4. In Section 5 we demonstrate that the extended approach covers all scenarios of Byrne's suppression task and present further results. In the final Section 6 we discuss our findings and suggest some future research.

2 The Suppression Task

Ruth Byrne (Byrne 1989) has conducted a number of experiments where subjects (not trained in logic) were asked to draw various conclusions given certain sets of sentences. In order to present the experiments in a compact form we will make use of the abbreviations shown in Table 1. Furthermore, $\neg X$ shall denote the negative fact corresponding to the fact X, i.e. $\neg e$ denotes that *she does not have an essay to write*.

Table 2 summarizes the results reported in (Byrne 1989). E.g., the third experiment (in comparison to the first one) shows that the addition of the sentence C_o to C_e, e leads to the suppression of l, although l is still entailed by C_o, C_e, e

C_e	*If she has an essay to write she will study late in the library.*
C_t	*If she has a textbook to read she will study late in the library.*
C_o	*If the library stays open she will study late in the library.*
e	*She has an essay to write.*
l	*She will study late in the library.*
o	*The library stays open.*
t	*She has textbooks to read.*

Table 1: Some abbreviations.

in case of a (naive) representation of the sentences in classical propositional logic.

The experiments have been repeated several times leading to similar figures (see e.g. (Dieussaert et al. 2000)).

3 A Logic for Human Reasoning

As mentioned in Section 1 the first step of the approach by Keith Stenning and Michiel van Lambalgen (Stenning and van Lambalgen 2008) consists of reasoning towards an appropriate logical representation of the sentences. As this step is not under consideration in this paper, we simply repeat their proposal without further discussion.

Keith Stenning and Michiel van Lambalgen consider logic programs, where the atoms occurring in the body of a clause can be either $\top$ (denoting the truth value *true*), $\bot$ (denoting the truth value *false*), or standard atoms. In particular, if A is an atom then $A \leftarrow \top$ denotes a *positive fact*, whereas $A \leftarrow \bot$ denotes a so-called *negative fact*. The latter becomes clear only if we apply a completion semantics (see below).

One of the main ideas in (Stenning and van Lambalgen 2008) is to represent conditionals by licences for conditionals using abnormality predicates. E.g., C_e, e is represented by the program

$$\mathcal{P}_{ee} \;=\; \{l \leftarrow e \wedge \neg ab,\; e \leftarrow \top,\; ab \leftarrow \bot\}.$$

Likewise, C_e, C_t, e and C_e, C_o, e are represented by

$$\mathcal{P}_{ete} \;=\; \{l \leftarrow e \wedge \neg ab_1,\; l \leftarrow t \wedge \neg ab_2,\; e \leftarrow \top,\; ab_1 \leftarrow \bot,\; ab_2 \leftarrow \bot\}$$

and

$$\mathcal{P}_{eoe} \;=\; \{l \leftarrow e \wedge \neg ab_1,\; l \leftarrow o \wedge \neg ab_2,\; e \leftarrow \top,\; ab_1 \leftarrow \neg o,\; ab_2 \leftarrow \neg e\},$$

respectively.

These programs are completed using a weak form of completion which is identical with Clark's completion (Clark 1978) except that undefined predicates stay undefined and are not declared to be false (see (Hölldobler and Ramli 2009b)). E.g., as the *weak completion* of the above mentioned programs we obtain:

$$\begin{array}{lcl} wc\,\mathcal{P}_{ee} & = & \{l \leftrightarrow e \wedge \neg ab,\; e \leftrightarrow \top,\; ab \leftrightarrow \bot\}, \\ wc\,\mathcal{P}_{ete} & = & \{l \leftrightarrow (e \wedge \neg ab_1) \vee (t \wedge \neg ab_2),\; e \leftrightarrow \top, \\ & & \quad ab_1 \leftrightarrow \bot,\; ab_2 \leftrightarrow \bot\}, \\ wc\,\mathcal{P}_{eoe} & = & \{l \leftrightarrow (e \wedge \neg ab_1) \vee (o \wedge \neg ab_2),\; e \leftrightarrow \top, \\ & & \quad ab_1 \leftrightarrow \neg o,\; ab_2 \leftrightarrow \neg e\}. \end{array}$$

Programs and their (weak) completions are evaluated by three-valued interpretations. Such interpretations are represented by tuples of the form $\langle I^\top, I^\bot \rangle$, where $I^\top$ denotes the set of all atoms which are mapped to *true*, $I^\bot$ denotes the set of all atoms which are mapped to *false*, $I^\top$ and $I^\bot$ are disjoint, and all atoms which do neither occur in $I^\top$ nor in $I^\bot$ are mapped to *undefined* or *unknown*.

If we choose the three-valued Łukasiewicz semantics (Łukasiewicz 1920) then logic programs enjoy the model intersection property, i.e., for each program, the intersection of all models is itself a model. Moreover, the model intersection property holds for weakly completed programs as well, and each model for the weak completion of a program is also a model for the program. See (Hölldobler and Ramli 2009b) for details. It should be noted that these properties do not hold if we consider the strong Kleene semantics with complete equivalence as done in (Stenning and van Lambalgen 2008).

The *least* model of a program is a model $\langle I^\top, I^\bot \rangle$ such that there does not exist another model $\langle J^\top, J^\bot \rangle$ with $J^\top \subset I^\top$ and $J^\bot \subseteq I^\bot$, or $J^\top \subseteq I^\top$ and $J^\bot \subset I^\bot$. It can be computed as the least fixed point of the following operator introduced in (Stenning and van Lambalgen 2008): Let I be an interpretation and $\mathcal{P}$ a program. Then, $\Phi_\mathcal{P}^{SvL} = \langle J^\top, J^\bot \rangle$, where

$$\begin{array}{lcl} J^\top & = & \{A \mid \text{there exists } A \leftarrow \mathit{body} \in \mathcal{P} \text{ with } I(\mathit{body}) = \mathit{true}\}, \\ J^\bot & = & \{A \mid \text{there exists } A \leftarrow \mathit{body} \in \mathcal{P} \text{ and for all } A \leftarrow \mathit{body} \in \mathcal{P} \text{ we find } I(\mathit{body}) = \mathit{false}\}. \end{array}$$

One should observe the subtle difference in the first line of the definition of $J^\bot$ if compared to the so-called *Fitting* operator usually associated with three-valued logic programs (see (Fitting 1985)).

As shown in (Hölldobler and Ramli 2009b; 2009c) the first six of Ruth Byrne's experiments (the first column in Table 2) are adequately modeled by considering the least model of corresponding weakly completed programs under Łukasiewicz semantics. For example, the least model of $wc\,\mathcal{P}_{eoe}$ is $\langle \{e\},\; \{ab_2\} \rangle$ from which we conclude that it is *unknown* whether she studies late in the library.

But what about the second column in Table 2? In order to model these experiments we need to add abduction to the framework presented so far.

K	Q	A	K	Q	A
C_e, e	l	96%	C_e, l	e	55%
C_e, C_t, e	l	96%	C_e, C_t, l	e	16%
C_e, C_o, e	l	38%	C_e, C_o, l	e	55%
$C_e, \neg e$	$\neg l$	46%	$C_e, \neg l$	$\neg e$	69%
$C_e, C_t, \neg e$	$\neg l$	4%	$C_e, C_t, \neg l$	$\neg e$	69%
$C_e, C_o, \neg e$	$\neg l$	63%	$C_e, C_o, \neg l$	$\neg e$	44%

Table 2: A brief summary of Ruth Byrne's experiments, where K denotes the given set of sentences, Q denotes the query and A denotes the percentage of positive answers.

4 Abduction

Let $\mathcal{L}$ be a language, $\mathcal{K} \subseteq \mathcal{L}$ a set of formulas called *knowledge base*, $\mathcal{A} \subseteq \mathcal{L}$ a set of formulas called *abducibles* and $\models \subseteq 2^{\mathcal{L}} \times \mathcal{L}$ a logical *consequence relation*. Following (Kakas, Kowalski, and Toni 1993), the triple $\langle \mathcal{K}, \mathcal{A}, \models \rangle$ is called an *abductive framework*. An *observation* $\mathcal{O}$ is a subset of $\mathcal{L}$; it is *explained* by $\mathcal{E}$ (or $\mathcal{E}$ is an *explanation* for $\mathcal{O}$) iff $\mathcal{E} \subseteq \mathcal{A}$, $\mathcal{K} \cup \mathcal{E}$ is satisfiable, and $\mathcal{K} \cup \mathcal{E} \models L$ for each $L \in \mathcal{O}$. An explanation $\mathcal{E}$ for $\mathcal{O}$ is said to be *minimal* iff there is no explanation $\mathcal{E}' \subset \mathcal{E}$ for $\mathcal{O}$.

Here we consider abductive frameworks that are instantiated in the following way: The knowledge base $\mathcal{K}$ is a logic program $\mathcal{P}$ where $\mathcal{L}$ is the language underlying $\mathcal{P}$. Let $\mathcal{R}_\mathcal{P}$ be the set of relation symbols occurring in $\mathcal{P}$, let

$$\mathcal{R}^D_\mathcal{P} = \{A \in \mathcal{R}_\mathcal{P} \mid A \leftarrow \mathit{body} \in \mathcal{P}\}$$

be the set of *defined relation symbols* in $\mathcal{P}$ and let $\mathcal{R}^U_\mathcal{P} = \mathcal{R}_\mathcal{P} \setminus \mathcal{R}^D_\mathcal{P}$ be the set of *undefined relation symbols* in $\mathcal{P}$. Then, the set of abducibles is

$$\mathcal{A} = \{A \leftarrow \top \mid A \in \mathcal{R}^U_\mathcal{P}\} \cup \{A \leftarrow \bot \mid A \in \mathcal{R}^U_\mathcal{P}\}.$$

The consequence relation $\models$ is $\models^{\text{lm wc}}_{3Ł}$, where $\mathcal{P} \models^{\text{lm wc}}_{3Ł} F$ iff F is mapped to true under the least model of the weak completion of $\mathcal{P}$ using the three-valued Łukasiewicz semantics. The observation $\mathcal{O}$ is usually a set containing a single literal L, in which case we simply write $\mathcal{O} = L$ instead of $\mathcal{O} = \{L\}$. A formula $F \in \mathcal{L}$ *follows sceptically by abduction* from $\mathcal{P}$ and $\mathcal{O}$, in symbols $\mathcal{P}, \mathcal{O} \models^s_A F$, iff $\mathcal{O}$ can be explained and for all minimal explanations $\mathcal{E}$ we find $\mathcal{P} \cup \mathcal{E} \models^{\text{lm wc}}_{3Ł} F$. A formula $F \in \mathcal{L}$ *follows credulously by abduction* from $\mathcal{P}$ and $\mathcal{O}$, in symbols $\mathcal{P}, \mathcal{O} \models^c_A F$, iff there exists a minimal explanation $\mathcal{E}$ for $\mathcal{O}$ such that $\mathcal{P} \cup \mathcal{E} \models^{\text{lm wc}}_{3Ł} F$.

5 Results

The Suppression Task Let us consider the experiments presented in the second column of Table 2. First, we will show that they can be adequately represented within the developed framework. To this end let

$$\begin{array}{lcl} \mathcal{P}_e & = & \{l \leftarrow e \wedge \neg ab,\ ab \leftarrow \bot\}, \\ \mathcal{P}_{et} & = & \{l \leftarrow e \wedge \neg ab_1,\ ab_1 \leftarrow \bot,\ l \leftarrow t \wedge \neg ab_2, \\ & & \ ab_2 \leftarrow \bot\}, \\ \mathcal{P}_{eo} & = & \{l \leftarrow e \wedge \neg ab_1,\ ab_1 \leftarrow \neg o,\ l \leftarrow o \wedge \neg ab_2, \\ & & \ ab_2 \leftarrow \neg e\} \end{array}$$

be the appropriate representation for C_e, C_e, C_t and C_e, C_o, respectively, obtained in the first step of the approach by Keith Stenning and Michiel van Lambalgen (Stenning and van Lambalgen 2008).

1. Consider $\mathcal{P}_e$ and let $\mathcal{O} = l$: $\mathcal{A} = \{e \leftarrow \top,\ e \leftarrow \bot\}$, $lm\, wc\, \mathcal{P}_e = \langle \emptyset, \{ab\}\rangle$, $\{e \leftarrow \top\}$ is the only minimal explanation for l, and $\mathcal{P}_e, l \models^s_A e$.
2. Consider $\mathcal{P}_{et}$ and let $\mathcal{O} = l$: $\mathcal{A} = \{e \leftarrow \top,\ e \leftarrow \bot,\ t \leftarrow \top,\ t \leftarrow \bot\}$, $lm\, wc\, \mathcal{P}_{et} = \langle \emptyset, \{ab_1,\ ab_2\}\rangle$, $\{e \leftarrow \top\}$ and $\{t \leftarrow \top\}$ are the minimal explanations for l, and $\mathcal{P}_{et}, l \not\models^s_A e$.
3. Consider $\mathcal{P}_{eo}$ and let $\mathcal{O} = l$: $\mathcal{A} = \{e \leftarrow \top,\ e \leftarrow \bot,\ o \leftarrow \top,\ o \leftarrow \bot\}$, $lm\, wc\, \mathcal{P}_{et} = \langle \emptyset, \emptyset\rangle$, $\{e \leftarrow \top,\ o \leftarrow \top\}$ is the only minimal explanation for l, and $\mathcal{P}_{eo}, l \models^s_A e$.
4. Consider $\mathcal{P}_e$ and let $\mathcal{O} = \neg l$: $\mathcal{A} = \{e \leftarrow \top,\ e \leftarrow \bot\}$, $lm\, wc\, \mathcal{P}_e = \langle \emptyset, \{ab\}\rangle$, $\{e \leftarrow \bot\}$ is the only minimal explanation for $\neg l$, and $\mathcal{P}_e, \neg l \models^s_A \neg e$.
5. Consider $\mathcal{P}_{et}$ and let $\mathcal{O} = \neg l$: $\mathcal{A} = \{e \leftarrow \top,\ e \leftarrow \bot,\ t \leftarrow \top,\ t \leftarrow \bot\}$, $lm\, wc\, \mathcal{P}_{et} = \langle \emptyset, \{ab_1,\ ab_2\}\rangle$, $\{e \leftarrow \bot,\ t \leftarrow \bot\}$ is the only minimal explanation for $\neg l$, and $\mathcal{P}_{et}, \neg l \models^s_A \neg e$.
6. Consider $\mathcal{P}_{eo}$ and let $\mathcal{O} = \neg l$: $\mathcal{A} = \{e \leftarrow \top,\ e \leftarrow \bot,\ o \leftarrow \top,\ o \leftarrow \bot\}$, $lm\, wc\, \mathcal{P}_{et} = \langle \emptyset, \emptyset\rangle$, $\{e \leftarrow \bot\}$ and $\{o \leftarrow \bot\}$ are minimal explanations for $\neg l$, and $\mathcal{P}_{eo}, \neg l \not\models^s_A \neg e$.

In other words, the formalization appears to be adequate with respect to the findings reported in (Byrne 1989).

Variations In this paragraph we discuss some examples which demonstrate that the various elements of the proposed formalization are needed. In (Hölldobler and Ramli 2009c; 2009b) it has already been shown that the strong three-valued Kleene logic with complete equivalence is inadequate to model all of the experiments mentioned in the first column of Table 2.

Reconsider the case of modus ponens with positive observation (case 1. above), but consider $\langle \mathcal{P}_e, \mathcal{A}, \models_{3Ł}\rangle$ instead of $\langle \mathcal{P}_e, \mathcal{A}, \models^{\text{lm wc}}_{3Ł}\rangle$, where $\models_{3Ł}$ is the usual entailment relation with respect to the three-valued Łukasiewicz logic. One should observe that in such a logic least models may not exist. In this case neither $\mathcal{P}_e \cup \{e \leftarrow \top\} \models_{3Ł} l$ nor $\mathcal{P}_e \cup \{e \leftarrow \bot\} \models_{3Ł} l$ because ab can be mapped to true. Hence, the observation l can not be explained at all (in contrast to (Byrne 1989)). The example demonstrates that weak completion is needed.

Consider the case of modus ponens with negative observation (case 4. above), but consider now $\langle \mathcal{P}_e, \mathcal{A}, \models^{\text{c}}_{3Ł}\rangle$ instead of $\langle \mathcal{P}_e, \mathcal{A}, \models^{\text{lm wc}}_{3Ł}\rangle$, where $\mathcal{P} \models^{\text{c}}_{3Ł} F$ iff F holds in all models for the completion of $\mathcal{P}$. The completion of $\mathcal{P}_e$ is $\{l \leftrightarrow e \wedge \neg ab,\ ab \leftrightarrow \bot,\ e \leftrightarrow \bot\}$, which entails $\neg l$, i.e. the empty set is an explanation. Hence, we find that $\mathcal{P}_e, \neg l \not\models^s_A \neg e$ (in contrast to (Byrne 1989)). The example demonstrates that completion is insufficient.

Reconsider again the case of modus ponens with negative observation (case 4. above), but weakly complete only the program $\mathcal{P}_e$ and not the explanation. In this case we find that neither $wc\, \mathcal{P}_e$ nor $wc\, \mathcal{P}_e \cup \{e \leftarrow \top\}$ nor $wc\, \mathcal{P}_e \cup \{e \leftarrow \bot\}$ nor $wc\, \mathcal{P}_e \cup \{e \leftarrow \top,\ e \leftarrow \bot\}$ entails $\neg l$. Hence, the observation l cannot be explained (in contrast to (Byrne 1989)). The example demonstrates that explanations must be (weakly) completed as well.

Reconsider the case of alternative arguments with positive observation (case 2. above), but now reason credulously instead of sceptically. There are two minimal explanations, viz. $\{e \leftarrow \top\}$ as well as $\{e \leftarrow \bot\}$. Hence, $\mathcal{P}_{et}, l \not\models^s_A e$, but $\mathcal{P}_{et}, l \models^c_A e$. Credulous reasoning is inconsistent with (Byrne 1989).

Extending Abduction by Integrity Constraints In this paragraph, we consider the application of abduction to human reasoning, where abductive explanations are restricted by integrity constraints (Kakas, Kowalski, and Toni 1993). We construct variants of some of the suppression task scenarios by Byrne and show that abduction with integrity constraints yields plausible results, suggesting that corresponding experiments should be made.

Here, an *integrity constraint* IC is a formula of the form $\bot \leftarrow (\neg)A_1 \wedge \ldots \wedge (\neg)A_n$. Integrity constraints are considered with two alternative semantics, the theoremhood view and the consistency view (Kakas, Kowalski, and Toni 1993). An *explanation satisfies IC in the theoremhood view* iff $\mathcal{E} \cup \mathcal{P} \models_{3Ł}^{\text{lm wc}} IC$. An *explanation satisfies IC in the consistency view* iff there exists an interpretation I such that $I_L \models wc\mathcal{E} \cup \mathcal{P} \cup \{IC\}$.

We extend the four scenarios C_e, C_t, l, C_e, C_o, l, $C_e, C_t, \neg l$, and $C_e, C_o, \neg l$ by Byrne with the following two phrases:

> *She will not read a textbook in holidays. There are holidays.*
>
> *The library is not open in holidays. There are holidays.*

The first sentence of each of the two phrases is encoded as a constraint, $IC_t = \bot \leftarrow t \wedge h$ and $IC_o = \bot \leftarrow o \wedge h$, respectively. The shared second sentence "There are holidays" is translated just as a fact $h \leftarrow \top$. We consider the following extension of the programs shown in the paragraph on the suppression task: $\mathcal{P}_{eth} = \mathcal{P}_{et} \cup \{h \leftarrow \top\}$ and $\mathcal{P}_{eoh} = \mathcal{P}_{eo} \cup \{h \leftarrow \top\}$.

1. Consider $\mathcal{P}_{eth}$ and let $\mathcal{O} = l$: $\mathcal{A} = \{e \leftarrow \top,\ e \leftarrow \bot,\ t \leftarrow \top,\ t \leftarrow \bot\}$, $lm\ wc\,\mathcal{P}_{eth} = \langle\{h\}, \{ab_1,\ ab_2\}\rangle$, $\{e \leftarrow \top, t \leftarrow \bot\}$ is the only explanation that satisfies IC_t in the theoremhood view; $\{e \leftarrow \top\}$ is the only minimal explanation for IC_t under the consistency view
2. Consider $\mathcal{P}_{eoh}$ and let $\mathcal{O} = l$: $\mathcal{A} = \{e \leftarrow \top,\ e \leftarrow \bot,\ o \leftarrow \top,\ o \leftarrow \bot\}$, $lm\ wc\,\mathcal{P}_{eoh} = \langle\{h\}, \emptyset\rangle$, $\{e \leftarrow \top,\ o \leftarrow \top\}$ is the only minimal explanation for l. However, this explanation does not satisfy IC_o neither in the consistency nor in the theoremhood view.
3. Consider $\mathcal{P}_{eth}$ and let $\mathcal{O} = \neg l$: $\mathcal{A} = \{e \leftarrow \top,\ e \leftarrow \bot,\ t \leftarrow \top,\ t \leftarrow \bot\}$, $lm\ wc\,\mathcal{P}_{et} = \langle\{h\}, \{ab_1,\ ab_2\}\rangle$, $\{e \leftarrow \bot,\ t \leftarrow \bot\}$ is the only minimal explanation for $\neg l$, which satisfies IC_t in the consistency as well as in the theoremhood view.
4. Consider $\mathcal{P}_{eoh}$ and let $\mathcal{O} = \neg l$: $\mathcal{A} = \{e \leftarrow \top,\ e \leftarrow \bot,\ o \leftarrow \top,\ o \leftarrow \bot\}$, $lm\ wc\,\mathcal{P}_{et} = \langle\emptyset, \emptyset\rangle$, $\{o \leftarrow \bot\}$ is the minimal explanation that satisfies IC_o in the consistency view, $\{o \leftarrow \bot\}$ also satisfies IC_o in the consistency view with the model $\langle\{ab_1, ab_2, h\}, \{e, l, o\}\rangle$ and $\{e \leftarrow \bot\}$ with the model $\langle\{ab_1, h\}, \{o, l\}\rangle$.

Consider the last case under the theoremhood view: Here, $\{e \leftarrow \bot\}$ is not an explanation. This is interesting since one reason that she is not in the library could be that she does not have an essay to write. In this sense, the theoremhood view eliminates meaningful explanations. This is not the case in the consistency view.

Consistency:	NP-*complete*
Relevance:	NP
Necessity:	coNP-*complete*
Skeptical Reasoning:	DP-*complete*

Table 3: Complexity classes of considered abductive tasks.

Complexity Results In this paragraph we discuss the complexity of four abductive tasks: (1.) consistency, i.e. the question whether there exists a minimal explanation, (2.) relevance, i.e. the question whether there exists a minimal explanation containing a specific fact (3.) necessity, i.e. whether all minimal explanations contain a specific fact and (4.) the complexity of sceptical reasoning. Table 3 shows the complexity classes of these problems.

We first take a close look at the consistency problem. Suppose, we already have a set $\mathcal{E}$ and we want to decide if it is a minimal explanation for an observation $\mathcal{O}$. It is easy to check whether $\mathcal{E} \subseteq \mathcal{A}$, and $\mathcal{K} \cup \mathcal{E} \models_{3Ł}^{\text{lm wc}} L$ for each $L \in \mathcal{O}$. In order to compute the least model one can use the least fixed point of $\Phi_{\mathcal{P}\cup\mathcal{E}}^{SvL}$, which can be computed in polynomial time. The condition that $\mathcal{K} \cup \mathcal{E}$ is satisfiable can be dropped since there always exists a least Łukasiewicz model of a weakly completed program. It remains to decide whether an explanation $\mathcal{E}$ is minimal. There are $2^{|\mathcal{E}|} - 1$ strict subsets which have to be checked whether they are explanations or nor. However, this exponential blowup can be avoided. In classical logic, minimality can be decided in polynomial time by iterating over all $F \in \mathcal{E}$ and testing whether $\mathcal{E} \setminus \{F\}$ is an explanation or not. If there is no such explanation, then $\mathcal{E}$ is minimal. Otherwise, $\mathcal{E}$ is not minimal. That this is correct follows from the fact that classical logic is monotonic (see (Hermann and Pichler 2007, Theorem 5)). We say a logic is *monotonic* iff $\mathcal{F} \models G$ implies $\mathcal{F} \cup \mathcal{F}' \models G$, for all sets of formulas $\mathcal{F}, \mathcal{F}'$ and formulas G. If we consider the least model of a weakly completed program under the Łukasiewicz semantics, then we do not have a monotonic logic: Consider the empty program $\mathcal{P}$ and $G = A \leftrightarrow C$. Then $lm\ wc\,\mathcal{P} = \langle\emptyset, \emptyset\rangle \models G$. By adding $A \leftarrow \top$ to $\mathcal{P}$, we have $lm\ wc\,\mathcal{P} \cup \{A \leftarrow \top\} = \langle\{A\}, \emptyset\rangle \not\models G$. However, in the considered abductive problems, we restrict $\mathcal{F}'$ to be a subset of

$$\{A \leftarrow \top \mid A \in \mathcal{R}_{\mathcal{P}}^{U}\} \cup \{A \leftarrow \bot \mid A \in \mathcal{R}_{\mathcal{P}}^{U}\}.$$

Then, the following holds:

If $\mathcal{E}$ is an explanation, then any non-contradictory extension of $\mathcal{E}$ is an explanation.

This result is surprising: Although the the consequence operator $\models_{3Ł}^{\text{lm wc}}$ is not monotonic, abductive explanations are monotonic. This means that one can safely extend an explanation by further non-contradictory facts. The reasons why we obtain this property are that we require that explanations cannot be further explained and the observation is a set of literals. Moreover, with this relation between the least model of the original program $\mathcal{P}$ and the extended program, we can decide minimality as follows: *An explanation $\mathcal{E}$ is minimal iff $\mathcal{E} \setminus \{f\}$ is not an explanation for all $f \in \mathcal{E}$.*

NP-membership of consistency can be shown follows: Guess a minimal explanation $\mathcal{E}$. To verify if $lm\ wc\ \mathcal{P} \cup \mathcal{E} \models L$ for all $L \in \mathcal{O}$ holds, compute the least fixed point of $\Phi^{SvL}_{\mathcal{P}\cup\mathcal{E}}$. Since this operator is monotonic, we obtain the least fixed point after polynomially many applications. Afterwards, we check whether $\mathcal{E} \setminus \{f\}$ is an explanation for all $f \in \mathcal{E}$ to verify minimality. NP-*hardness* follows by a reduction from 3SAT. Consider the following transformation: Let $F = C_1 \wedge \ldots \wedge C_n$ be a 3SAT instance and $X_1 \ldots X_m$ the variables occurring in F. Then, the abductive problem is obtained as follows:

$$\begin{array}{rcl} AF & = & \langle \mathcal{P}, \{X_i \leftarrow \top, X_i \leftarrow \bot \mid 1 \leq i \leq m\}\rangle \\ \mathcal{O} & = & \{O\} \\ \mathcal{P} & = & \{Y_i \leftarrow L_{i,1}, Y_i \leftarrow L_{i,2}, Y_i \leftarrow L_{i,3} \mid \\ & & \quad \text{for each clause } C_i = L_{i,1} \vee L_{i,2} \vee L_{i,3}\} \\ & & \cup\, \{O \leftarrow Y_1 \wedge \ldots \wedge Y_n\} \end{array}$$

Then the following holds: F is satisfiable iff there exists a minimal explanation for $\mathcal{O}$. The reason why this is correct is that one can easily construct an explanation from a model of F and vice versa. 3SAT is known to be NP-*hard* and since polynomial time reductions are transitive, we can conclude that consistency is also NP-*hard*. It immediately follows that consistency is NP-*complete* and inconsistency is coNP-*complete*.

It is easy to see that the second considered problem, relevance, is not harder than consistency: One has simply to guess a minimal explanation containing a specific fact.

Necessity and inconsistency are equivalent w.r.t. polynomial time reductions, which can be shown as follows: Let $\langle \mathcal{P}, \mathcal{A}, \models\rangle$ be an abductive framework and $\mathcal{O}$ an observation. Suppose, we want to decide if f is necessary in every explanation for $\mathcal{O}$. Then, this problem is equivalent to the question whether $\mathcal{O}$ is not explainable in $\langle \mathcal{P}, \mathcal{A} \setminus \{f\}, \models^{\text{lm wc}}_{3Ł}\rangle$, i.e. it is inconsistent. Suppose we want to decide whether there does not exists a solution at all. Then, this problem is equivalent to the question whether q is necessary in $\langle \mathcal{P}, \mathcal{A} \cup \{q \leftarrow \top, q \leftarrow \bot\}\rangle$ where q is a fresh atom. Since inconsistency is coNP-*complete*, we obtain that necessity is coNP-*complete*.

The fourth considered problem is skeptical reasoning. Consider the class DP: A language L belongs to the class DP, if there are two languages L_1, L_2 such that $L = L_1 \cap L_2$, L_1 belongs to NP and L_2 belongs to coNP. Sceptical reasoning consists of two sub problems, where consistency is already shown to be NP-*complete*. Consider the complement of the second problem, i.e. does there exists a minimal explanation $\mathcal{E}$ with $\mathcal{P} \cup \mathcal{E} \not\models^{\text{lm wc}}_{3Ł} F$? It is clear that this problem is in NP, since one have to simply guess the correct minimal explanations and minimality can be checked in polynomial time. Hence, the original problem is in coNP. coNP-*hardness* follows by a reduction from necessity: A fact $A \leftarrow \top$ ($A \leftarrow \bot$) is necessary iff for all minimal explanations $\mathcal{E}$ we find that $\mathcal{P} \cup \mathcal{E} \models^{\text{lm wc}}_{3Ł} A$ ($\mathcal{P} \cup \mathcal{E} \models^{\text{lm wc}}_{3Ł} \neg A$) . DP-*hardness* follows immediately by the fact that both problems are hard. Hence, sceptical reasoning is DP-*complete*.

6 Conclusion

Logic appears to be adequate for human reasoning if weak completion, the three-valued Łukasiewicz semantics, the semantic operator $\Phi^{SvL}_{\mathcal{P}}$, and abduction are used. Human reasoning is modeled by, firstly, reasoning towards an appropriate logic program $\mathcal{P}$ and, secondly, by reasoning with respect to the least model of the weak completion of the $\mathcal{P}$ (which is equal to the least fixed point of $\Phi^{SvL}_{\mathcal{P}}$) and, in case of abduction, by taking a sceptical point of view. This approach matches data from studies in human reasoning and, in particular, the data first reported in (Byrne 1989). However, much remains to be done.

There is a connectionist encoding of the approach (Hölldobler and Ramli 2009c) which, unfortunately, does not yet include abduction. On the other hand, various proposals to handle abduction in a connectionist setting have been made (e.g. (d'Avila Garcez et al. 2007)); these proposals are more or less straightforward encodings of a sequential search in the space of all possible explanations and they model only credulous reasoning. How do humans search for explanations? In which order are explanations generated by humans if there are several? Do humans prefer minimal explanations? Does attention play a role in the selection of explanations? Do humans reason sceptically or credulously? How does a connectionist realization of abductive reasoning embedded into (Hölldobler and Ramli 2009c) looks like?

In a Łukasiewicz logic the semantic deduction theorem does not hold. Is this adequate with respect to human reasoning? Likewise, in the three-valued Łukasiewicz logic an implication is mapped to *true* if both, its precondition and conclusion, are mapped to *unknown*. How do humans evaluate implications whose precondition and conclusion mapped to *unknown*?

In the current approach negative and positive facts are not treated on the same level. Rather, by considering the weak completion of a program negative facts are dominated by positive information. How is negation treated in human reasoning?

In (Hölldobler and Ramli 2009a) it was shown that the semantic operator $\Phi^{SvL}_{\mathcal{P}}$ associated with a program $\mathcal{P}$ (see Section 3) is a contraction if $\mathcal{P}$ is acyclic. In this case, thanks to Banach's contraction mapping theorem, $\Phi^{SvL}_{\mathcal{P}}$ admits a unique fixed point which can be computed by iterating $\Phi^{SvL}_{\mathcal{P}}$ starting with an arbitrary initial interpretation. Do humans exhibit a behaviour which can be adequately modeled by contractional semantic operators? If so, can we generate appropriate level mappings (needed to show acyclicity of a program) by studying the behavior of humans?

Last but not least, what is the relation between the proposed approach and well-founded and/or stable and/or circumscription-projection (Wernhard 2010) semantics?

References

Byrne, R. 1989. Suppressing valid inferences with conditionals. *Cognition* 31:61–83.

Clark, K. 1978. Negation as failure. In Gallaire, H., and Minker, J., eds., *Logic and Databases*. New York: Plenum. 293–322.

d'Avila Garcez, A.; Gabbay, D.; O.Ray; and Woods, J. 2007. Abductive reasoning in neural-symbolic learning systems. *TOPOI* 26:37–49.

Dieussaert, K.; Schaeken, W.; Schroyen, W.; and d'Ydevalle, G. 2000. Strategies during complex conditional inferences. *Thinking and Reasoning* 6(2):152–161.

Evans, J. S. T.; Newstead, S. E.; and Byrne, R. M. J. 1993. *Human Reasoning – The Psychology of Deduction*. Lawrence Erlbaum Associates.

Fitting, M. 1985. A Kripke–Kleene semantics for logic programs. *Journal of Logic Programming* 2(4):295–312.

Hermann, M., and Pichler, R. 2007. Counting complexity of propositional abduction. In *In Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI 07)*, 417–422.

Hölldobler, S., and Ramli, C. K. 2009a. Contraction properties of a semantic operator for human reasoning. In Li, L., and Yen, K. K., eds., *Proceedings of the Fifth International Conference on Information*, 228–231. International Information Institute.

Hölldobler, S., and Ramli, C. K. 2009b. Logic programs under three-valued Łukasiewicz's semantics. In Hill, P., and Warren, D., eds., *Logic Programming*, volume 5649 of *LNCS*, 464–478. Springer Berlin Heidelberg.

Hölldobler, S., and Ramli, C. K. 2009c. Logics and networks for human reasoning. In et. al., C. A., ed., *ICANN*, volume 5769 of *LNCS*, 85–94. Springer Berlin Heidelberg.

Kakas, A. C.; Kowalski, R. A.; and Toni, F. 1993. Abductive Logic Programming. *Journal of Logic and Computation* 2(6):719–770.

Kleene, S. 1952. *Introduction to Metamathematics*. North-Holland.

Łukasiewicz, J. 1920. O logice trójwartościowej. *Ruch Filozoficzny* 5:169–171. English translation: On Three-Valued Logic. In: *Jan Łukasiewicz Selected Works*. (L. Borkowski, ed.), North Holland, 87-88, 1990.

McCarthy, J. 1963. Situations and actions and causal laws. Stanford Artificial Intelligence Project: Memo 2.

Stenning, K., and van Lambalgen, M. 2008. *Human Reasoning and Cognitive Science*. MIT Press.

Wernhard, C. 2010. Circumscription and projection as primitives of logic programming. In *Technical Communications of the 26th International Conference on Logic Programming, ICLP'10*, volume 7 of *Leibniz International Proceedings in Informatics (LIPIcs)*.