A Computational Logic Approach to the Belief Bias in Human Syllogistic Reasoning

Emmanuelle-Anna Dietz(✉)

International Center for Computational Logic, TU Dresden, Dresden, Germany
emmanuelle.dietz@tu-dresden.de

Abstract. Psychological experiments on syllogistic reasoning have shown that participants did not always deduce the classical logically valid conclusions. In particular, the results show that they had difficulties to reason with syllogistic statements that contradicted their own beliefs. We consider a syllogistic reasoning task carried out by Evans, Barston and Pollard, who investigated the belief-bias effect with respect to syllogisms. We propose a formalization of the belief-bias effect for human syllogistic reasoning under the Weak Completion Semantics, a logic programming approach that aims at adequately modeling human reasoning.

1 Introduction

Evans et al. [15] carried out a psychological study about deductive reasoning, which demonstrated possibly conflicting processes in human reasoning. Participants were presented different syllogisms for which they had to decide whether they accepted these syllogisms as valid. Consider S_{vit}:

Premise 1	*No nutritional things are inexpensive.*
Premise 2	*Some vitamin tablets are inexpensive.*
Conclusion	*Some vitamin tablets are not nutritional.*

The Conclusion necessarily follows from the premises under classical logic. However, about half of the participants said that the syllogism was not valid. They were explicitly asked to logically validate or invalidate various syllogisms, but did not seem to have the intellectual capability to do so. Even worse, they were not even aware about their inabilities. Participants reflectively read the instructions and understood well that they were required to reason logically from the premises to the conclusion. However, the results show that their intuitions were stronger and delivered a tendency to say '*yes*' or '*no*' depending on whether the syllogism was believable [14]. The responses of participants for various syllogisms, which differed with respect to their validity and whether they were believable in the contextual setting, were evaluated in [15]. Four of them are depicted in Table 1. The first two premises of all four cases are of the same logical form, namely *No A are B. Some C are B.* The first two cases differ from the last two cases with respect to the conclusions: In the first two cases the conclusions correspond to the logical form *Some C are A* and in the last two

P. Brézillon et al. (Eds.): CONTEXT 2017, LNAI 10257, pp. 691–707, 2017.
DOI: 10.1007/978-3-319-57837-8_55

Table 1. Four types of syllogisms taken from [11]. The percentages of the participants that accepted the syllogism as being valid are shown in the last column.

	Type	Case	%
S_{dog}	Valid and believable	*No police dogs are vicious*	92
		Some highly trained dogs are vicious	
		Some highly trained dogs are not police dogs	
S_{vit}	Valid and unbelievable	*No nutritional things are inexpensive*	46
		Some vitamin tablets are inexpensive	
		Some vitamin tablets are not nutritional	
S_{rich}	Invalid and unbelievable	*No millionaires are hard workers*	8
		Some rich people are hard workers	
		Some millionaires are not rich people	
S_{cig}	Invalid and believable	*No addictive things are inexpensive*	92
		Some cigarettes are inexpensive	
		Some addictive things are not cigarettes	

cases the conclusions correspond to the logical form *Some A are C*. The first two syllogisms are indeed valid under classical logic, whereas the last two are not. However, as the last column shows, the percentage of the participants that validated the syllogism, does not necessarily comply with the results under classical logic. Evans, Barston and Pollard asserted that the participants were influenced by their own beliefs, their so-called belief bias.

Khemlani and Johnson-Laird [24] have compared the predictions of 12 cognitive theories to participants' responses in syllogistic reasoning. The Verbal Model Theory [28] performed best with an accurate prediction of 84%, closely followed by the Mental Model Theory [21], which achieved 83%. Recently, [3] developed a logical form for the representation of syllogisms under the logic programming approach, the Weak Completion Semantics, and predicted even 89% of the participants' responses.

The Weak Completion Semantics is a new cognitive theory, which originates from [30], but is mathematically sound [19], and has been successfully applied – among others – to the suppression task [8], the selection task [9] to reasoning about conditionals [5,7] and to spatial reasoning [6]. As the Weak Completion Semantics aims at modeling human reasoning adequately and predicted well the participants' responses in syllogistic reasoning, a natural question to ask, is whether the belief-bias effect in syllogistic reasoning can be modeled within this approach.

After briefly discussing the belief-bias effect, we introduce the Weak Completion Semantics. Taking [3,4] as starting point, Sects. 4 and 5 present six principles for modeling quantified statements in human reasoning and show their representations in logic programs. Finally, Sect. 6 presents how the belief-bias effect can be modeled under the Weak Completion Semantics by discussing the four cases in Table 1.

2 The Belief-Bias Effect

Evans et al. [13] distinguish between the negative and the positive belief bias: The negative belief bias describes the case when the support for an unbelievable conclusion is suppressed. On the other hand, the positive belief bias describes the case when the acceptance for a believable conclusion is raised. Consider again Table 1: The negative belief bias happens for 46% of the participants in the case of S_{vit} and the positive belief bias happens for 92% of the participants in the case of S_{cig}.

As pointed out in [16], Wilkins [31] already observed that syllogisms which conflict with our beliefs are more difficult to solve. Since then, various theories have tried to explain why humans deviate from the classical logically valid answers. Some conclusions can be explained by converting the premises as proposed in [2] or by assuming that the type of the premises creates an atmosphere which influences the acceptance for the conclusion [16,32]. Johnson-Laird and Byrne [22] proposed the mental model theory [21], which additionally supposes the search for counterexamples when validating the conclusion. Later, Stenning and van Lambalgen [30] explain why certain aspects influence the interpretations made by humans when evaluating syllogisms and discuss this in the context of mental models. Evans et al. [10,15] proposed a theory, which in the literature is sometimes referred to as the selective scrutiny model [1,16]. First, humans heuristically accept any syllogism having a believable conclusion, and only proceed with a logical evaluation if the conclusion contradicts their belief. Adler and Rips [1] claim that this behavior is rational in the sense of efficient belief maintenance. Yet another approach, the selective processing model [12], accounts only for a single preferred model: If the conclusion is neutral or believable, humans attempt to construct a model that supports it. Otherwise, they attempt to construct a model that rejects it.

According to Garnham and Oakhill [16] the belief-bias effect can take place at several stages: First, beliefs can influence our understanding of the premises. Second, in case a statement contradicts our belief, we might search for alternative models and check whether the conclusion is plausible. This seems to comply with Stenning and van Lambalgen's proposal to model human reasoning by a two step procedure [30]. The first step, the representational part, determines how our beliefs influence the understanding of the premises. The second step, the procedural part, determines whether we search for alternative models based on the plausibility of the conclusion.

In this paper we will follow up on this distinction when modeling the belief-bias effect.

Table 2. $\top$, $\bot$, and U denote *true*, *false*, and *unknown*, respectively.

F	$\neg F$
$\top$	$\bot$
$\bot$	$\top$
U	U

$\wedge$	$\top$	U	$\bot$
$\top$	$\top$	U	$\bot$
U	U	U	$\bot$
$\bot$	$\bot$	$\bot$	$\bot$

$\vee$	$\top$	U	$\bot$
$\top$	$\top$	$\top$	$\top$
U	$\top$	U	U
$\bot$	$\top$	U	$\bot$

$\leftarrow$	$\top$	U	$\bot$
$\top$	$\top$	$\top$	$\top$
U	U	$\top$	$\top$
$\bot$	$\bot$	U	$\top$

$\leftrightarrow$	$\top$	U	$\bot$
$\top$	$\top$	U	$\bot$
U	U	$\top$	U
$\bot$	$\bot$	U	$\top$

3 Weak Completion Semantics

The general notation, which we will use in the paper, is based on [18,25].

3.1 Logic Programs

We restrict ourselves to datalog programs, i.e., the set of terms consists only of constants and variables.

$$A \leftarrow L_1 \wedge \ldots\ldots \wedge L_n. \tag{1}$$
$$A \leftarrow \top. \tag{2}$$
$$A \leftarrow \bot. \tag{3}$$

A is an atom and the L_i with $1 \leq i \leq n$ are literals. The atom A is called *head* of the clause and the subformula to the right of the implication symbol is called *body* of the clause. If the clause contains variables, then they are implicitly universally quantified within the scope of the entire clause. A *ground clause* is a clause not containing variables. Clauses of the form (2) and (3) are called *facts* and *assumptions*, respectively. The notion of falsehood appears counterintuitive at first sight, but programs will be interpreted under their (weak) completion where we replace the implication by the equivalence sign. We assume a fixed set of constants, denoted by $\mathcal{C}$, which is nonempty and finite. $\mathsf{constants}(\mathcal{P})$ denotes the set of all constants occurring in $\mathcal{P}$. If not stated otherwise, we assume that $\mathcal{C} = \mathsf{constants}(\mathcal{P})$. $g\mathcal{P}$ denotes ground $\mathcal{P}$, which means that $\mathcal{P}$ contains exactly all the ground clauses with respect to the alphabet. $\mathsf{atoms}(\mathcal{P})$ denotes the set of all atoms occurring in $g\mathcal{P}$. If atom A is not the head of any clause in $\mathcal{P}$, then A is *undefined* in $g\mathcal{P}$. The set of all atoms that are undefined in $g\mathcal{P}$ is $\mathsf{undef}(\mathcal{P})$.

3.2 Three-Valued Łukasiewicz Logic

We consider the three-valued Łukasiewicz logic [26], for which the corresponding truth values are $\top$, $\bot$ and U, which mean *true*, *false* and *unknown*, respectively. A *three-valued interpretation* I is a mapping from formulas to the set of truth values $\{\top, \bot, \mathrm{U}\}$. The truth value of a given formula under I is determined according to the truth tables in Table 2. We represent an interpretation as a pair $I = \langle I^\top, I^\bot \rangle$ of disjoint sets of atoms where $I^\top$ is the set of all atoms that are mapped to $\top$ by I, and $I^\bot$ is the set of all atoms that are mapped to $\bot$

by I. Atoms, which do not occur in $I^\top \cup I^\bot$, are mapped to U. Let $I = \langle I^\top, I^\bot \rangle$ and $J = \langle J^\top, J^\bot \rangle$ be two interpretations: $I \subseteq J$ iff $I^\top \subseteq J^\top$ and $I^\bot \subseteq J^\bot$. $I(F) = \top$ means that a formula F is mapped to true under I. $\mathcal{M}$ is a *model* of $g\mathcal{P}$ if it is an interpretation, which maps each clause occurring in $g\mathcal{P}$ to $\top$. I is the *least model* of $g\mathcal{P}$ iff for any other model J of $g\mathcal{P}$ it holds that $I \subseteq J$.

3.3 Reasoning with Respect to Least Models

Consider the following transformation for $\mathcal{P}$: 1. Replace all clauses in $g\mathcal{P}$ with the same head $A \leftarrow body_1, A \leftarrow body_2, \ldots$ by the single expression $A \leftarrow body_1 \vee body_2, \vee \ldots$. 2. Replace all occurrences of $\leftarrow$ by $\leftrightarrow$. The resulting set of equivalences is called the *weak completion* of $\mathcal{P}$ ($\mathsf{wc}\,\mathcal{P}$). The model intersection property holds for weakly completed programs, which guarantees the existence of a least model for every $\mathcal{P}$ [20]. Stenning and van Lambalgen [30] devised the following operator, which has been generalized for first-order programs in [19]: Let I be an interpretation in $\Phi_\mathcal{P}(I) = \langle J^\top, J^\bot \rangle$, where

$J^\top = \{A \mid$ there exists $A \leftarrow body \in g\mathcal{P}$ and $I(body) = \top\}$,
$J^\bot = \{A \mid A \notin \mathsf{undef}(\mathcal{P})$ and for all $A \leftarrow body \in g\mathcal{P}$ we find that $I(body) = \bot\}$.

As shown in [19] the least fixed point of $\Phi_\mathcal{P}$ is identical to the least model of the weak completion of $g\mathcal{P}$ under three-valued Łukasiewicz logic ($\mathsf{lm}\,\mathsf{wc}\,\mathcal{P}$). Starting with $I = \langle \emptyset, \emptyset \rangle$, $\mathsf{lm}\,\mathsf{wc}\,\mathcal{P}$ is computed by iterating $\Phi_\mathcal{P}$. Given a program $\mathcal{P}$ and a formula F, $\mathcal{P} \models_{wcs} F$ iff $\mathsf{lm}\,\mathsf{wc}\,\mathcal{P}(F) = \top$ for formula F.

3.4 Integrity Constraints

A set of *integrity constraints* $\mathcal{IC}$ comprises clauses of the form $\mathrm{U} \leftarrow body$, where *body* is a conjunction of literals. Given $\mathcal{P}$ and $\mathcal{IC}$, $\mathcal{P}$ *satisfies* $\mathcal{IC}$ iff for all $\mathrm{U} \leftarrow body \in \mathcal{IC}$, we find that $\mathcal{P} \models_{wcs} \mathrm{U} \leftarrow body$ (i.e. $\mathcal{P} \not\models_{wcs} body$).

3.5 Abduction

We extend two-valued abduction [23] for three-valued semantics. The set of abducibles $\mathcal{A}_\mathcal{P}$ may not only contain facts but can also contain assumptions:

$$\mathcal{A}_\mathcal{P} = \{A \leftarrow \top \mid A \in \mathsf{undef}(\mathcal{P})\} \cup \{A \leftarrow \bot \mid A \in \mathsf{undef}(\mathcal{P})\}.$$

Let $\langle \mathcal{P}, \mathcal{A}_\mathcal{P}, \mathcal{IC}, \models_{wcs} \rangle$ be an abductive framework, $\mathcal{E} \subset \mathcal{A}_\mathcal{P}$ and observation $\mathcal{O}$ a non-empty set of literals.

$\mathcal{O}$ is *explained by* $\mathcal{E}$ *given* $\mathcal{P}$ *and* $\mathcal{IC}$ iff $\mathcal{P} \cup \mathcal{E} \models_{wcs} \mathcal{O}$ and $\mathcal{P} \cup \mathcal{E} \models_{wcs} \mathcal{IC}$.
$\mathcal{O}$ is *explained given* $\mathcal{P}$ and $\mathcal{IC}$ iff there exists $\mathcal{E}$ s.t. $\mathcal{O}$ is explained by $\mathcal{E}$ given $\mathcal{P}$ and $\mathcal{IC}$.

We assume that explanations are minimal, i.e. there is no other explanation $\mathcal{E}' \subset \mathcal{E}$ for $\mathcal{O}$. We distinguish between skeptical and credulous reasoning in abduction as follows:

- F *follows skeptically from* $\mathcal{P}, \mathcal{IC}$ and $\mathcal{O}$ iff $\mathcal{O}$ can be explained given $\mathcal{P}$ and $\mathcal{IC}$, and for all minimal $\mathcal{E}$ for $\mathcal{O}$ given $\mathcal{P}$ and $\mathcal{IC}$, it holds that $\mathcal{P} \cup \mathcal{E} \models_{wcs} F$.
- F *follows credulously from* $\mathcal{P}$, $\mathcal{IC}$ *and* $\mathcal{O}$ iff there exists a minimal $\mathcal{E}$ for $\mathcal{O}$ given $\mathcal{P}$ and $\mathcal{IC}$, and it holds that $\mathcal{P} \cup \mathcal{E} \models_{wcs} F$.

In the following, we are interested in deriving skeptically entailed information. The entailment relation $\models^s_{wcs}$ is an abbreviation to express that a formula follows skeptically, i.e. $\mathcal{P}, \mathcal{IC}, \mathcal{O} \models^s_{wcs} F$ denotes that F follows skeptically from $\mathcal{P}, \mathcal{IC}$ and $\mathcal{O}$.

4 Six Principles on Quantified Statements

We introduce six principles for developing the representation of quantified statements and reasoning with respect to them, originally developed in [3]. Some are motivated by ideas from the area of Logic Programming and others are motivated by findings from Cognitive Science.

4.1 Licenses for Inferences (lice)

Stenning and van Lambalgen [30] propose to formalize conditionals in human reasoning not by inferences straight away, but rather by *licenses for inferences*. For instance, the conditional '*if* $y(X)$ *then* $z(X)$' is represented by the program, which consists of

$$z(X) \leftarrow y(X) \wedge \neg ab_{yz}(X). \qquad ab_{yz}(X) \leftarrow \bot.$$

The first clause states that '$z(X)$ *if* $y(X)$ *and* $\neg ab_{yz}(X)$'. The second clause represents the closed-world assumption with respect to $ab_{yz}(X)$, where $ab_{yz}(X)$ is an abnormality predicate. We call this principle *licenses for inferences* (lice).

4.2 Negation by Transformation (trans)

The logic programs we consider under the Weak Completion Semantics do not allow heads of clauses to be negative literals. In order to represent a negative conclusion $\neg y(X)$, we introduce an auxiliary formula $y'(X)$ together with the clause $y(X) \leftarrow \neg y'(X)$ and the integrity constraint $\mathrm{U} \leftarrow y(X) \wedge y'(X)$. This is a widely used technique in logic programming. Together with the principle introduced in Sect. 4.1 (lice), this additional clause is extended by the following two clauses:

$$y(X) \leftarrow \neg y'(X) \wedge \neg ab_{nyy}(X). \qquad ab_{nyy}(X) \leftarrow \bot.$$

Additionally, the integrity constraint $\mathrm{U} \leftarrow y(X) \wedge y'(X)$ states that an object cannot belong to both, y and y'. We call this principle *negation by transformation* (trans).

4.3 Existential Import and Gricean Implicature (import)

Normally, we do not quantify over objects that do not exist. Accordingly, '*all y are z*' implies '*some y are z*', which is referred to as *existential import* and implied by *Gricean implicature* [17]. Existential import is assumed by the theory of mental models [21] or mental logic [29]. Likewise, humans require existential import for a conditional to be true [30]. Furthermore, the quantifier '*some y are z*' often implies that '*some y are not z*', which again is implied by the Gricean implicature [24]: Someone would not state '*some y are z*' if that person knew that '*all y are z*'. As the person does not say '*all y are z*' but '*some y are z*', we assume that '*not all y are z*', which in turn implies '*some y are not z*'. We call this principle *existential import and Gricean implicature* (import).

4.4 Unknown Generalization (unkGen)

Humans seem to distinguish between '*some y are z*' and '*some z are y*' [24]. However, if we would represent '*some y are z*' by $\exists X(y(X) \wedge z(X))$ then this is semantically equivalent to $\exists X(z(X) \wedge y(X))$ because conjunction is commutative in first-order logic. Likewise, as we have discussed in Sect. 4.3, humans seem to distinguish between '*some y are z*' and '*all y are z*'. Accordingly, if we only observe that an object belongs to y and z then we do not want to conclude both, '*some y are z*' and '*all y are z*'. Therefore we introduce the following principle: If we know that '*some y are z*' then there must not only be an object, which belongs to y and z (by Gricean implicature) but there must be another object, which belongs to y and for which it is unknown whether it belongs to z. We call this principle *unknown generalization* (unkGen).

4.5 No Derivation Through Double Negation (dNeg)

Under the Weak Completion Semantics, a positive conclusion can be derived from double negation within two conditionals. Consider the following two conditionals with each having a negative premise: If not x, then y. If not y then z. Additionally, assume that x is true. Let $\mathcal{P} = \{z \leftarrow \neg y,\ y \leftarrow \neg x,\ x \leftarrow \top\}$ be the program that encodes this information. The $\mathsf{lm}\,\mathsf{wc}\,\mathcal{P}$ is $\langle\{x, z\}, \{y\}\rangle$: x is true because it is a fact and y is false because the negation of x is false. z is true by the negation of y. However, considering the results in [24], humans seem not to draw conclusions through double negatives. Accordingly, we block them with the abnormalities introduced by principle (import) in Sect. 4.1. We call this principle *no derivation through double negation* (dNeg).

4.6 Search for Alternative Models (searchAlt)

Consider again S_{rich} and S_{add}: The premises are about things which contradict the conclusion. We assume that in case there seems no conclusion possible, humans might try to search for alternative models by explaining some part of the information that is presented. We call this principle *Search for Alternative Models* (searchAlt).

5 Representation of Quantified Statements as Programs

Based on the first five principles of the previous section, we encode the quantified statements in logic programs, where y and z will later be replaced by the properties of the corresponding objects. Note that, different to the principles in Sects. 4.1 to 4.5, principle (searchAlt) in Sect. 4.6 is not about the representation of the quantified statements but about the reasoning process, which will be discussed later. Note that the capital letters in brackets in the title of each of the following subsections, A, E, I and O are the classical abbreviations for the quantifiers *All*, *No*, *Some* and *Some not*.

5.1 All y are z (Ayz)

'*All y are z*' is represented by the program $\mathcal{P}_{Ayz}$, which consists of the following clauses:

$$z(X) \leftarrow y(X) \wedge \neg ab_{yz}(X). \quad \text{(lice)}$$
$$ab_{yz}(X) \leftarrow \bot. \quad \text{(lice)}$$
$$y(o) \leftarrow \top. \quad \text{(import)}$$

The least model of the weak completion of $\mathcal{P}_{Ayz}$, $\mathsf{lm\,wc}\,\mathcal{P}_{Iyz}$, is $\langle\{y(o), z(o)\}, \{ab_{yz}(o)\}\rangle$.

5.2 No y is z (Eyz)

Under FOL 'No y is z' is represented as $\forall X(y(X) \rightarrow \neg z(X))$, which is equivalent to $\forall X(z(X) \rightarrow \neg y(X))$. $\mathcal{P}_{Eyz}$ consists of the following clauses:

$$y'(X) \leftarrow z(X) \wedge \neg ab_{zny}(X). \quad \text{(trans \& lice)}$$
$$ab_{zny}(X) \leftarrow \bot. \quad \text{(lice)}$$
$$y(X) \leftarrow \neg y'(X) \wedge \neg ab_{nyy}(X). \quad \text{(trans \& lice)}$$
$$z(o) \leftarrow \top. \quad \text{(import)}$$
$$ab_{nyy}(o) \leftarrow \bot. \quad \text{(lice \& dNeg)}$$

We have the following integrity constraint: $\mathrm{U} \leftarrow y(X) \wedge y'(X)$. (trans)
Note that the last clause in $\mathcal{P}_{Eyz}$ cannot be generalized to all X, because otherwise we allow conclusions by double negatives: principle (dNeg) states that we should block conclusions through double negatives. The least model of the weak completion of $\mathcal{P}_{Eyz}$, $\mathsf{lm\,wc}\,\mathcal{P}_{Eyz}$, is $\langle\{y'(o), z(o)\}, \{ab_{zny}(o), ab_{nyy}(o), y(o)\}\rangle$.

5.3 Some y are z (Iyz)

'*Some y are z*' is represented by the program $\mathcal{P}_{Iyz}$:

$$z(X) \leftarrow y(X) \wedge \neg ab_{yz}(X). \quad \text{(lice)}$$
$$ab_{yz}(o_1) \leftarrow \bot. \quad \text{(unkGen \& lice)}$$
$$y(o_1) \leftarrow \top. \quad \text{(import)}$$
$$y(o_2) \leftarrow \top. \quad \text{(unkGen)}$$

$\mathsf{lm\,wc}\,\mathcal{P}_{Iyz}$ is $\langle\{y(o_1), y(o_2), z(o_1)\}, \{ab_{yz}(o_1)\}\rangle$.
Nothing about $ab_{yz}(o_2)$ is stated in $\mathcal{P}_{Iyz}$. Accordingly, $z(o_2)$ stays unknown in $\mathsf{lm\,wc}\,\mathcal{P}_{Iyz}$.

5.4 Some y are Not z (Oyz)

'*Some y are not z*' is represented by the program $\mathcal{P}_{Oyz}$:

$$
\begin{array}{rll}
z'(X) \leftarrow y(X) \wedge \neg ab_{ynz}(X). & & (\mathsf{trans}\ \&\ \mathsf{lice}) \\
ab_{ynz}(o_1) \leftarrow \bot. & & (\mathsf{unkGen}\ \&\ \mathsf{lice}) \\
z(X) \leftarrow \neg z'(X) \wedge \neg ab_{nzz}(X). & & (\mathsf{trans}\ \&\ \mathsf{lice}) \\
y(o_1) \leftarrow \top. & & (\mathsf{import}) \\
y(o_2) \leftarrow \top. & & (\mathsf{unkGen}) \\
ab_{nzz}(o_1) \leftarrow \bot. & & (\mathsf{dNeg}\ \&\ \mathsf{lice}) \\
ab_{nzz}(o_2) \leftarrow \bot. & & (\mathsf{dNeg}\ \&\ \mathsf{lice})
\end{array}
$$

We have the following integrity constraint: $U \leftarrow z(X) \wedge z'(X)$. (trans)
The first four clauses as well as the integrity constraint are derived as in the program $\mathcal{P}_{Eyz}$ except that object o_1 is used instead of o and ab_{ynz} is restricted to o_1 as in $\mathcal{P}_{Iyz}$. The fifth clause of $\mathcal{P}_{Oyz}$ is obtained by principle (unkGen). The last two clauses are not generalized to all objects for the same reason as discussed in Sect. 5.2: The generalization of ab_{nzz} to all objects would lead to conclusions through double negation in case there would be a second premise. The least model of the weak completion of $\mathcal{P}_{Oyz}$, $\mathsf{lm\,wc}\,\mathcal{P}_{Oyz}$, is $\langle\{y(o_1), y(o_2), z'(o_1)\}, \{ab_{ynz}(o_1), ab_{nzz}(o_1), ab_{nzz}(o_2), z(o_1)\}\rangle$.

5.5 Entailment of the Quantified Statements

We specify when Ayz, Eyz, Iyz or Oyz are entailed by a model.

- $\mathcal{P} \models$ Ayz iff there exists an object o such that $\mathcal{P} \models_{wcs} y(o)$ and for all objects o we find that if $\mathcal{P} \models_{wcs} y(o)$ then $\mathcal{P} \models_{wcs} z(o)$.
- $\mathcal{P} \models$ Eyz iff there exists an object o such that $\mathcal{P} \models_{wcs} z(o)$ and for all objects o we find that if $\mathcal{P} \models_{wcs} z(o)$ then $\mathcal{P} \models_{wcs} \neg y(o)$.
- $\mathcal{P} \models$ Iyz iff there exists an object o_1 such that $\mathcal{P} \models_{wcs} y(o_1) \wedge z(o_1)$ and there exists an object o_2 such that $\mathcal{P} \models_{wcs} y(o_2)$ and $\mathcal{P} \not\models_{wcs} z(o_2)$.
- $\mathcal{P} \models$ Oyz iff there exists an object o_1 such that $\mathcal{P} \models_{wcs} y(o_1) \wedge \neg z(o_1)$ and there exists an object o_2 such that $\mathcal{P} \models_{wcs} y(o_2)$ and $\mathcal{P} \not\models_{wcs} \neg z(o_2)$.

If nothing can be concluded, i.e. if $\mathcal{P} \not\models$ Ayz, $\mathcal{P} \not\models$ Eyz, $\mathcal{P} \not\models$ Iyz and $\mathcal{P} \not\models$ Oyz, then principle (searchAlt) applies and we search for alternative models by trying to explain y. Later, y refers to the first property in the conclusion of the syllogism and z refers to the two properties left. For instance, consider S_{dog}: y refers to *highly trained dogs* (*high_trai*) and z is either *police dogs* (*pol_dog*) or *vicious* (*vic*). If nothing between either *high_trai* and *pol_dog* or between *high_trai* and *vic* can be derived, we try to explain *high_trai*.

6 Modeling the Belief-Bias Effect

According to the observations made in Sect. 2, we model the belief-bias effect in two stages: (1) the belief can influence the representation, i.e. how the given information is understood, or (2) the belief can influence the reasoning, i.e. how new information is gained, if nothing can be derived. In the following, we model (1) with help of abnormalities, motivated by principle (lice). (2) is modeled by means of skeptical abduction, motivated by principle (searchAlt). The following four syllogisms are modeled according to the logic program representations proposed in Sects. 5.2 and 5.4.

6.1 No Belief-Bias Effect

$\mathcal{P}_{dog}$ represents the first two premises of S_{dog} and consists of

$$
\begin{aligned}
pol\_dog'(X) &\leftarrow vic(X) \wedge \neg ab_{pol\_dog'}(X). && \text{(trans \& lice)}\\
ab_{pol\_dog'}(X) &\leftarrow \bot. && \text{(lice)}\\
pol\_dog(X) &\leftarrow \neg pol\_dog'(X) \wedge \neg ab_{pol\_dog}(X). && \text{(trans \& lice)}\\
vic(o_1) &\leftarrow \top. && \text{(import)}\\
ab_{pol\_dog}(o_1) &\leftarrow \bot. && \text{(lice \& dNeg)}\\
vic(X) &\leftarrow high\_trai(X) \wedge \neg ab_{vic}(X). && \text{(lice)}\\
ab_{vic}(o_2) &\leftarrow \bot. && \text{(unkGen \& lice)}\\
high\_trai(o_2) &\leftarrow \top. && \text{(import)}\\
high\_trai(o_3) &\leftarrow \top. && \text{(unkGen)}
\end{aligned}
$$

We have the following integrity constraint: $\mathrm{U} \leftarrow pol\_dog(X) \wedge pol\_dog'(X)$. (trans)
$\mathsf{lm\,wc}\,\mathcal{P}_{dog} = \langle I^{\top}, I^{\bot} \rangle$, is as follows:

$$
\begin{aligned}
I^{\top} &= \{high\_trai(o_2), high\_trai(o_3), pol\_dog'(o_1), pol\_dog'(o_2), vic(o_1), vic(o_2)\},\\
I^{\bot} &= \{pol\_dog(o_2), pol\_dog(o_1), ab_{pol\_dog'}(o_1), ab_{pol\_dog'}(o_2), ab_{pol\_dog'}(o_3),\\
&\quad\ ab_{pol\_dog}(o_1), ab_{vic}(o_2)\},
\end{aligned}
$$

Indeed, this model entails the Conclusion of S_{dog}, *Some highly trained dogs are not police dogs*: There exists an object, o_2, such that $\mathcal{P}_{dog} \models_{wcs} high\_trai(o_2) \wedge \neg pol\_dog(o_2)$ and there exists another object, o_3, such that $\mathcal{P}_{dog} \models_{wcs} high\_trai(o_3)$ and $\mathcal{P}_{dog} \not\models_{wcs} \neg pol\_dog(o_3)$. According to [15], S_{dog} is logically valid and psychologically believable. No conflict arises either at the psychological or at the logical level. The majority validated the syllogism, which complies with what is entailed by $\mathsf{lm\,wc}\,\mathcal{P}_{dog}$.

6.2 Belief-Bias Effect in Representation Stage

$\mathcal{P}_{vit}$ represents the first two premises of S_{vit} and consists of

$$
\begin{array}{rcll}
nutri'(X) & \leftarrow & inex(X) \wedge \neg ab_{nutri'}(X). & (\mathsf{trans}\ \&\ \mathsf{lice}) \\
ab_{nutri'}(X) & \leftarrow & \bot. & (\mathsf{lice}) \\
nutri(X) & \leftarrow & \neg nutri'(X) \wedge \neg ab_{nutri}(X). & (\mathsf{trans}\ \&\ \mathsf{lice}) \\
inex(o_1) & \leftarrow & \top. & (\mathsf{import}) \\
ab_{nutri}(o_1) & \leftarrow & \bot. & (\mathsf{lice}\ \&\ \mathsf{dNeg}) \\
inex(X) & \leftarrow & vitamin(X), \neg ab_{inex}(X). & (\mathsf{lice}) \\
ab_{inex}(o_2) & \leftarrow & \bot. & (\mathsf{unkGen}\ \&\ \mathsf{lice}) \\
vitamin(o_2) & \leftarrow & \top. & (\mathsf{import}) \\
vitamin(o_3) & \leftarrow & \top. & (\mathsf{unkGen})
\end{array}
$$

We have the following integrity constraint: $\mathrm{U} \leftarrow nutri'(X) \wedge nutri(X)$. (trans)

The corresponding $\mathsf{lm}\,\mathsf{wc}\,\mathcal{P}_{vit} = \langle I^\top, I^\bot \rangle$, is as follows:

$$
\begin{aligned}
I^\top &= \{vitamin(o_2), vitamin(o_3), inex(o_1), inex(o_2), nutri'(o_1), nutri'(o_2)\} \\
I^\bot &= \{nutri(o_1), nutri(o_2), ab_{inex}(o_2), ab_{nutri}(o_1), ab_{nutri'}(o_1), ab_{nutri'}(o_2), \\
&\quad ab_{nutri'}(o_3)\},
\end{aligned}
$$

Indeed this model entails the Conclusion of S_{vit}, that *Some vitamin tablets are not nutritional*: There exists an object, o_2, such that $\mathcal{P}_{vit} \models_{wcs} vitamin(o_2) \wedge \neg nutri(o_2)$ and there exists another object, o_3, such that $\mathcal{P}_{vit} \models_{wcs} vitamin(o_3)$ and $\mathcal{P}_{vit} \not\models_{wcs} \neg nutri(o_3)$. The results of the psychological study in Table 1 indicate that there seemed to be two groups of participants: The group that validated the syllogism was not influenced by the bias with respect to nutritional things. Their understanding of the syllogism is reflected by $\mathcal{P}_{vit}$ and their conclusion complies with what is entailed by $\mathsf{lm}\,\mathsf{wc}\,\mathcal{P}_{vit}$. The participants who chose to invalidate the syllogism belong to the other group that has apparently been influenced by their belief. The belief bias occurred in the representation stage. Accordingly, we model this aspect with help of abnormality predicates as follows: Regarding both premises, it is commonly known that

The purpose of vitamin tablets is to aid nutrition.

This belief in the context of Premise 1 leads to

If something is a vitamin tablet, then it is abnormal. (*regarding* Premise 1 *of* S_{vit})

We extend $\mathcal{P}_{vit}$ accordingly, which results in

$$\mathcal{P}^{\mathsf{bias}}_{vit} = \mathcal{P}_{vit} \cup \{ab_{nutri'}(X) \leftarrow vitamin(X)\}.$$

Observe that $ab_{nutri'}(X) \leftarrow vitamin(X)$ overrides $ab_{nutri'}(X) \leftarrow \bot$ under the weak completion of $\mathcal{P}^{\mathsf{bias}}_{vit}$. $\mathsf{lm}\,\mathsf{wc}\,\mathcal{P}^{\mathsf{bias}}_{vit} = \langle I^\top, I^\bot \rangle$ is

$$
\begin{aligned}
I^\top &= \{inex(o_1), inex(o_2), vitamin(o_2), vitamin(o_3), ab_{nutri'}(o_2), ab_{nutri'}(o_3)\}, \\
I^\bot &= \{nutri'(o_2), nutri'(o_3), ab_{nutri}(o_1), ab_{inex}(o_2)\}.
\end{aligned}
$$

In this case, the Conclusion of S_{vit}, that *Some vitamin tablets are not nutritional*, is not entailed. Actually, nothing is stated about the relation between vitamin tablets and them (not) being nutritional. However, as trivally *Some vitamin tablets are inexpensive* holds, principle (searchAlt) does not apply and we are done. According to [15], S_{vit} is logically valid but psychologically unbelievable. There arises a conflict at the psychological level because we generally assume that the purpose of vitamin tablets is to aid nutrition. The participants who have been influenced by this belief did not validate the syllogism, which complies to the result above, as the Conclusion is not entailed by $\mathsf{lm}\,\mathsf{wc}\,\mathcal{P}_{vit}^{\mathsf{bias}}$ either.

6.3 Belief-Bias Effect in Reasoning Stage

$\mathcal{P}_{rich}$ represents the first two premises of S_{rich} and consists of

$$
\begin{aligned}
mil'(X) &\leftarrow hard\_wor(X) \wedge \neg ab_{mil'}(X). && (\mathsf{trans}\ \&\ \mathsf{lice})\\
ab_{mil'}(X) &\leftarrow \bot. && (\mathsf{lice})\\
mil(X) &\leftarrow \neg mil'(X) \wedge ab_{mil}(X). && (\mathsf{trans}\ \&\ \mathsf{lice})\\
hard\_wor(o_1) &\leftarrow \top. && (\mathsf{import})\\
ab_{mil}(o_1) &\leftarrow \bot. && (\mathsf{lice}\ \&\ \mathsf{dNeg})\\
hard\_wor(X) &\leftarrow rich(X) \wedge \neg ab_{hard\_wor}(X). && (\mathsf{lice})\\
ab_{hard\_wor}(o_2) &\leftarrow \bot. && (\mathsf{unkGen}\ \&\ \mathsf{lice})\\
rich(o_2) &\leftarrow \top. && (\mathsf{import})\\
rich(o_3) &\leftarrow \top. && (\mathsf{unkGen})
\end{aligned}
$$

We have the following integrity constraint: $\mathrm{U} \leftarrow mil(X) \wedge mil'(X).$ (trans)
Its least model of the weak completion, $\langle I^{\top}, I^{\bot}\rangle$, is as follows:

$$
\begin{aligned}
I^{\top} &= \{hard\_wor(o_1), hard\_wor(o_2), mil'(o_1), mil'(o_2), rich(o_2), rich(o_3)\},\\
I^{\bot} &= \{mil(o_1), mil(o_2), ab_{hard\_wor}(o_2), ab_{mil}(o_1), ab_{mil'}(o_1), ab_{mil'}(o_2),\\
&\quad\ ab_{mil'}(o_3)\},
\end{aligned}
$$

and does not confirm the Conclusion of S_{rich}, that *some millionaires are not rich people*. Actually, the Conclusion in S_{rich} states something which contradicts Premise 2, and cannot be about any of the previously introduced constant o_1, o_2 or o_3. As nothing can be derived about the relation between *mil* and *hard_wor* nor between *mil* and *rich*, principle (searchAlt) applies: According to our background knowledge, we know that 'normal' millionaires exist, i.e. millionaires for whom we do not assume anything abnormal with respect to them being millionaires. Additionally, we cannot be sure that all millionaires are normal, i.e. we know that millionaires exist for whom we simply don't know whether they are normal. We formulate this as an observation about two newly introduced constants, let's say o_4, representing a normal millionaire,[1] and o_5, representing a millionaire for whom it is unknown whether he or she is normal:

$$\mathcal{O} = \{mil(o_4), \neg ab_{mil'}(o_4), \neg ab_{mil}(o_4), mil(o_5)\}.$$

[1] This implies that all abnormalities about *mil* or *mil'* are false with respect to o_4.

If we want to find an explanation for $\mathcal{O}$ with respect to $\mathcal{P}_{mil}$, we can no longer assume that $\mathcal{C} = \mathsf{constants}(\mathcal{P}_{mil})$, because $\mathcal{A}_{\mathcal{P}_{mil}}$ does not contain any facts or assumptions about o_4 and o_5, as o_4 and o_5 do not occur in $\mathcal{P}_{mil}$. Therefore, we specify that the new set of constants under consideration is $\mathcal{C} = \{o_1, o_2, o_3, o_4, o_5\}$. Given that $\mathsf{lm\,wc}(\mathcal{P}_{mil}) = \langle I^\top, I^\bot \rangle$ as defined above, $\mathsf{lm\,wc}(\mathcal{P}^{\mathcal{C}}_{mil}) = \langle I^\top, I^\bot \cup \{ab_{mil'}(o_4), ab_{mil'}(o_5)\}\rangle$. The set of abducibles, $\mathcal{A}_{\mathcal{P}^{\mathcal{C}}_{mil}}$, contains six facts and six assumptions about o_4 and o_5:

$$\begin{array}{lll} rich(o_4) \leftarrow \top. & ab_{mil}(o_4) \leftarrow \top. & ab_{hard\_wor}(o_4) \leftarrow \top. \\ rich(o_4) \leftarrow \bot. & ab_{mil}(o_4) \leftarrow \bot. & ab_{hard\_wor}(o_4) \leftarrow \bot. \\ rich(o_5) \leftarrow \top. & ab_{mil}(o_5) \leftarrow \top. & ab_{hard\_wor}(o_5) \leftarrow \top. \\ rich(o_5) \leftarrow \bot. & ab_{mil}(o_5) \leftarrow \bot. & ab_{hard\_wor}(o_5) \leftarrow \bot. \end{array}$$

We find six (minimal) explanations for $\mathcal{O} = \{mil(o_4), \neg ab_{mil'}(o_4), \neg ab_{mil}(o_4), mil(o_5)\}$, where there are three from which the Conclusion of S_{rich} does not follow. Consider one of them, $\mathcal{E} = \{ab_{hard\_wor}(o_4) \leftarrow \top, ab_{mil}(o_4) \leftarrow \bot, ab_{mil}(o_5) \leftarrow \top\}$: Given that $\mathsf{lm\,wc}(\mathcal{P}_{mil}) = \langle I^\top, I^\bot \rangle$, $\mathsf{lm\,wc}(\mathcal{P}_{mil} \cup \mathcal{E})^{\mathcal{C}} = \langle J^\top, J^\bot \rangle$ is

$$\begin{aligned} J^\top &= I^\top \cup \{ab_{hard\_wor}(o_4), mil(o_4), mil(o_5), ab_{mil}(o_5)\}, \\ J^\bot &= I^\bot \cup \{ab_{mil}(o_4), ab_{mil'}(o_4), hard\_wor(o_4), mil'(o_4), ab_{mil'}(o_5)\}. \end{aligned}$$

According to the definition for skeptical abduction in Sect. 3.5, one explanation for which the Conclusion of S_{rich}, *Some millionaires are not rich people*, does not follow, is enough to show that the Conclusion does not follow skeptically from $\mathcal{P}^{\mathcal{C}}_{mil}$, $\mathcal{IC}$ and $\mathcal{O}$. According to [15] this case is neither logically valid nor believable. Almost no one validated S_{rich}, which complies to the result above, as the Conclusion is not skeptically entailed by $\mathcal{P}^{\mathcal{C}}_{mil}$, $\mathcal{IC}$ and $\mathcal{O}$ either.

6.4 Belief-Bias Effect in Representation and Reasoning Stage

$\mathcal{P}_{cig}$ represents the first two premises of S_{cig} and consists of

$$\begin{array}{rr} add'(X) \leftarrow inex(X) \wedge \neg ab_{add'}(X). & (\mathsf{trans}\ \&\ \mathsf{lice}) \\ ab_{add'}(X) \leftarrow \bot. & (\mathsf{lice}) \\ add(X) \leftarrow \neg add'(X) \wedge \neg ab_{add}(X). & (\mathsf{trans}\ \&\ \mathsf{lice}) \\ inex(o_1) \leftarrow \top. & (\mathsf{import}) \\ ab_{add}(o_1) \leftarrow \bot. & (\mathsf{lice}\ \&\ \mathsf{dNeg}) \\ inex(X) \leftarrow cig(X) \wedge \neg ab_{inex}(X). & (\mathsf{lice}) \\ ab_{inex}(o_2) \leftarrow \bot. & (\mathsf{unkGen}\ \&\ \mathsf{lice}) \\ cig(o_2) \leftarrow \top. & (\mathsf{import}) \\ cig(o_3) \leftarrow \top. & (\mathsf{unkGen}) \end{array}$$

We have the following integrity constraint: $\mathrm{U} \leftarrow add(X) \wedge add'(X)$. (trans)
It is commonly known that *Cigarettes are addictive*. This belief in the context of Premise 1 leads to

If something is a cigarette, then it is abnormal. (regarding Premise 1 *of* S_{cig}*)*

$\mathcal{P}_{cig}$ is extended accordingly. The new program is

$$\mathcal{P}_{cig}^{\mathsf{bias}} = \mathcal{P}_{cig} \cup \{ab_{add'}(X) \leftarrow cig(X)\}.$$

Observe that $ab_{add'}(X) \leftarrow cig(X)$ overrides $ab_{add'}(X) \leftarrow \bot$ under the weak completion of $\mathcal{P}_{cig}^{\mathsf{bias}}$. The least model of the weak completion of $\mathcal{P}_{cig}^{\mathsf{bias}}$, $\mathsf{lm\,wc}\,\mathcal{P}_{cig}^{\mathsf{bias}}$, is

$$\langle\{cig(o_2), cig(o_3), inex(o_1), inex(o_2)\}, \{ab_{add}(o_1), ab_{inex}(o_2)\}\rangle.$$

Similarly to the previous syllogism, this model does not state anything about the Conclusion, that *some addictive things are not cigarettes*. Again, the Conclusion of S_{cig} is about something, which cannot be o_1, o_2 or o_3. As nothing can be derived about the relation between *add* and *inex* nor between *add* and *cig*, principle (searchAlt) applies: According to our background knowledge, we know that 'normal' addictive things exist, i.e. addictive things for which we do not assume anything abnormal with respect to them being addictive things. Additionally, we cannot be sure that all addictive things are normal, i.e. we know that addictive things exist for which we simply don't know whether they are normal. We formulate this as an observation about two newly introduced constants, let's say o_4, representing normal addictive things[2] and o_5 representing addictive things for which it is unknown whether they are normal:

$$\mathcal{O} = \{add(o_4), \neg ab_{add'}(o_4), \neg ab_{add}(o_4), add(o_5)\}.$$

Let us define $\mathcal{C} = \{o_1, o_2, o_3, o_4, o_5\}$. $\mathsf{lm\,wc}\,\mathcal{P}_{cig}^{\mathsf{bias},\mathcal{C}}$ does not state anything about o_4 nor o_5: All atoms about o_4 and o_5 are unknown in this least model. Given $\mathcal{P}_{cig}^{\mathsf{bias},\mathcal{C}}$, the set of abducibles, $\mathcal{A}_{\mathcal{P}_{cig}^{\mathsf{bias},\mathcal{C}}}$ contains six facts and six assumptions about o_4 and o_5:

$$\begin{array}{lll} cig(o_4) \leftarrow \top. & ab_{add}(o_4) \leftarrow \top. & ab_{inex}(o_4) \leftarrow \top. \\ cig(o_4) \leftarrow \bot. & ab_{add}(o_4) \leftarrow \bot. & ab_{inex}(o_4) \leftarrow \bot. \\ cig(o_5) \leftarrow \top. & ab_{add}(o_5) \leftarrow \top. & ab_{inex}(o_5) \leftarrow \top. \\ cig(o_5) \leftarrow \bot. & ab_{add}(o_5) \leftarrow \bot. & ab_{inex}(o_5) \leftarrow \bot. \end{array}$$

The only three (minimal) explanations are

$$\mathcal{E}_1 = \mathcal{E}' \cup \{cig(o_5) \leftarrow \bot\},\ \mathcal{E}_2 = \mathcal{E}' \cup \{ab_{inex}(o_5) \leftarrow \bot\},\ \text{and}\ \mathcal{E}_3 = \mathcal{E}' \cup \{cig(o_5) \leftarrow \top\},$$

where $\mathcal{E}' = \{cig(o_4) \leftarrow \bot,\ ab_{add}(o_4) \leftarrow \bot,\ ab_{add}(o_5) \leftarrow \bot\}$. Given that $\mathsf{lm\,wc}\,(\mathcal{P}_{cig}^{\mathsf{bias}}) = \langle I^{\top}, I^{\bot}\rangle$ as defined above, the least models of the weak completion of $\mathcal{P}_{cig}^{\mathsf{bias},\mathcal{C}}$ together with the corresponding explanations, are as follows:

[2] This implies that all abnormalities about *add* or *add'* are false with respect to o_4.

$$\begin{aligned}
\mathsf{Im\,wc}\,(\mathcal{P}^{\mathsf{bias},\mathcal{C}}_{cig} \cup \mathcal{E}_1) &= \langle I^{\top} \cup \{add(o_4), add(o_5)\},\\
&\quad I^{\bot} \cup \{cig(o_4), inex(o_4), ab_{add}(o_4), ab_{add'}(o_4), add'(o_4),\\
&\quad ab_{add}(o_5), cig(o_5), ab_{add'}(o_5), inex(o_5), add'(o_5)\}\rangle,\\
\mathsf{Im\,wc}\,(\mathcal{P}^{\mathsf{bias},\mathcal{C}}_{cig} \cup \mathcal{E}_2) &= \langle I^{\top} \cup \{add(o_4), add(o_5), ab_{inex}(o_5)\},\\
&\quad I^{\bot} \cup \{cig(o_4), inex(o_4), ab_{add}(o_4), ab_{add'}(o_4), add'(o_4),\\
&\quad ab_{add}(o_5), inex(o_5), add'(o_5)\}\rangle,\\
\mathsf{Im\,wc}\,(\mathcal{P}^{\mathsf{bias},\mathcal{C}}_{cig} \cup \mathcal{E}_3) &= \langle I^{\top} \cup \{add(o_4), add(o_5), cig(o_5), ab_{add'}(o_5)\},\\
&\quad I^{\bot} \cup \{cig(o_4), inex(o_4), ab_{add}(o_4), ab_{add'}(o_4), add'(o_4),\\
&\quad ab_{add}(o_5), add'(o_5)\}\rangle.
\end{aligned}$$

The Conclusion of S_{add}, *Some addictive things are not cigarettes*, follows skeptically from $\mathcal{P}^{\mathsf{bias},\mathcal{C}}_{add}$ and $\mathcal{O}$, as the following derivation follows from all explanations for $\mathcal{O}$: There exists an object, o_4, such that $\mathcal{P}^{\mathsf{bias},\mathcal{C}}_{cig}, \mathcal{O} \models^{s}_{wcs} add(o_4) \wedge \neg cig(o_4)$ and there exists another object, o_5, such that $\mathcal{P}^{\mathsf{bias},\mathcal{C}}_{cig}, \mathcal{O} \models^{s}_{wcs} add(o_5)$ and $\mathcal{P}^{\mathsf{bias},\mathcal{C}}_{cig}, \mathcal{O} \not\models^{s}_{wcs} cig(o_5)$. According to [15], S_{cig} is classical logically invalid but psychologically believable and therefore causes a conflict: People are biased and search for a model that confirms their beliefs. This complies with what is entailed skeptically by $\mathcal{P}^{\mathsf{bias},\mathcal{C}}_{cig}$, $\mathcal{IC}$ and $\mathcal{O}$.

Note that in this case we need the restriction that explanations are minimal, otherwise $\mathcal{E}' \cup \{ab_{inex}(o_4) \leftarrow \top,\ cig(o_5) \leftarrow \bot\} \supset \mathcal{E}_1$ would be an explanation for $\mathcal{O}$ as well, and we could not derive that the Conclusion of S_{add} follows skeptically anymore.

7 Conclusion

By taking the principles presented in [3] as starting point and extending them with the additional principle *Search for Alternative Models*, we show how the belief-bias effect can be modeled by discussing the four cases of Evans et al.'s [15] syllogistic reasoning task. The belief-bias effect can be modeled in two stages: The first stage is where the belief bias seems to occur in the representational part of the syllogism, for instance in S_{vit}. In this case, the belief bias can be modeled by means of abnormality predicates. The belief bias in S_{cig} seems to occur in the representational and the reasoning part of the syllogism. The reasoning part can be modeled with skeptical abduction. Additionally, as the last case shows, explanations are required to be minimal.

To the best of our knowledge, the syllogistic reasoning tasks discussed in the literature have never accounted for providing the option 'I don't know' to the participants. As has been discussed in [27], participants who say that no valid conclusion follows, might have problems to actually find a conclusion easily, possibly meaning that they do not know the answer. The authors also point to [28], who suggested that if a conclusion is stated as being not valid this can mean that the reasoning process is exhausted. An experimental study, which allows the participants to distinguish between *I don't know* and *not valid*, might give us more insights about their reasoning processes.

Acknowledgements. Many thanks to Steffen Hölldobler and Luís Moniz Pereira for valuable feedback.

References

1. Adler, J., Rips, L.: Reasoning: Studies of Human Inference and Its Foundations. Cambridge University Press, Cambridge (2008)
2. Chapman, L.J., Chapman, J.P.: Atmosphere effect re-examined. J. Exp. Psychol. **58**(3), 220–226 (1959)
3. Costa, A., Dietz, E.A., Hölldobler, S., Ragni, M.: A computational logic approach to human syllogistic reasoning (2017, submitted)
4. Dietz, E.A.: A computational logic approach to syllogisms in human reasoning. In: Furbach, U., Schon, C. (eds.) CEUR WS Proceedings on Bridging the Gap Between Human and Automated Reasoning, pp. 17–31 (2015)
5. Dietz, E.-A., Hölldobler, S.: A new computational logic approach to reason with conditionals. In: Calimeri, F., Ianni, G., Truszczynski, M. (eds.) LPNMR 2015. LNCS (LNAI), vol. 9345, pp. 265–278. Springer, Cham (2015). doi:10.1007/978-3-319-23264-5_23
6. Dietz, E.A., Hölldobler, S., Höps, R.: A computational logic approach to human spatial reasoning. In: IEEE Symposium on Human-Like Intelligence (CIHLI) (2015)
7. Dietz, E.A., Hölldobler, S., Pereira, L.M.: On conditionals. In: Gottlob, G., Sutcliffe, G., Voronkov, A. (eds.) GCAI. Epic Series in Computing. EasyChair (2015)
8. Dietz, E.A., Hölldobler, S., Ragni, M.: A computational logic approach to the suppression task. In: Miyake, N., Peebles, D., Cooper, R.P. (eds.) Proceedings of 34th Conference of Cognitive Science Society, pp. 1500–1505. Cognitive Science Society (2012)
9. Dietz, E.A., Hölldobler, S., Ragni, M.: A computational logic approach to the abstract and the social case of the selection task. In: Proceedings of COMMONSENSE 2013 (2013)
10. Evans, J.S.: Bias in Human Reasoning - Causes and Consequences. Essays in Cognitive Psychology. Lawrence Erlbaum, Hove (1989)
11. Evans, J.S.: In two minds: dual-process accounts of reasoning. Trends Cogn. Sci. **7**(10), 454–459 (2003)
12. Evans, J.: Thinking and believing. In: Mental Models in Reasoning (2000)
13. Evans, J., Handley, S., Harper, C.: Necessity, possibility and belief: a study of syllogistic reasoning. Q. J. Exp. Psychol. **54**(3), 935–958 (2001)
14. Evans, J.: Biases in deductive reasoning. In: Pohl, R. (ed.) Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory. Psychology Press, New York (2012)
15. Evans, J., Barston, J.L., Pollard, P.: On the conflict between logic and belief in syllogistic reasoning. Memory Cogn. **11**(3), 295–306 (1983)
16. Garnham, A., Oakhill, J.: Thinking and Reasoning. Wiley, Hoboken (1994)
17. Grice, H.P.: Logic and conversation. In: Cole, P., Morgan, J.L. (eds.) Syntax and Semantics, vol. 3. Academic Press, New York (1975)
18. Hölldobler, S.: Logik und Logikprogrammierung 1: Grundlagen. Kolleg Synchron, Synchron (2009)
19. Hölldobler, S., Kencana Ramli, C.D.P.: Logic programs under three-valued Łukasiewicz semantics. In: Hill, P.M., Warren, D.S. (eds.) ICLP 2009. LNCS, vol. 5649, pp. 464–478. Springer, Heidelberg (2009). doi:10.1007/978-3-642-02846-5_37

20. Hölldobler, S., Kencana Ramli, C.D.P.: Logics and networks for human reasoning. In: Alippi, C., Polycarpou, M., Panayiotou, C., Ellinas, G. (eds.) ICANN 2009. LNCS, vol. 5769, pp. 85–94. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04277-5_9
21. Johnson-Laird, P.N.: Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness. Harvard University Press, Cambridge (1983)
22. Johnson-Laird, P.N., Byrne, R.M.: Deduction (1991)
23. Kakas, A.C., Kowalski, R.A., Toni, F.: Abductive logic programming. J. Log. Comput. **2**(6), 719–770 (1993)
24. Khemlani, S., Johnson-Laird, P.N.: Theories of the syllogism: a meta-analysis. Psychol. Bull. **138**, 427–457 (2012)
25. Lloyd, J.W.: Foundations of Logic Programming. Springer, New York (1984)
26. Łukasiewicz, J.: O logice trójwartościowej. In: Ruch Filozoficzny, vol. 5, pp. 169–171 (1920). English translation: On three-valued logic. In: Łukasiewicz, J., Borkowski, L. (eds.) Selected Works, pp. 87–88. North Holland, Amsterdam (1990)
27. Newstead, S., Handley, S., Buck, E.: Falsifying mental models: testing the predictions of theories of syllogistic reasoning. Memory Cogn. **27**(2), 344–354 (1999)
28. Polk, T.A., Newell, A.: Deduction as verbal reasoning. Psychol. Rev. **102**(3), 533–566 (1995)
29. Rips, L.J.: The Psychology of Proof: Deductive Reasoning in Human Thinking. MIT Press, Cambridge (1994)
30. Stenning, K., van Lambalgen, M.: Human Reasoning and Cognitive Science. A Bradford Book. MIT Press, Cambridge (2008)
31. Wilkins, M.: The effect of changed material on the ability to do formal syllogistic reasoning. Arch. Psychol. **16**(102), 1–83 (1928)
32. Woodworth, R.S., Sells, S.B.: An atmosphere effect in formal syllogistic reasoning. J. Exp. Psychol. **18**(4), 451–60 (1935)