Non-Monotonic Logics and Reasoning Biases

Catarina Dutilh Novaes
ILLC and Department of Philosophy
University of Amsterdam

Introduction

- Stenning and van Lambalgen (2008) advocate the usefulness of non-monotonic logics as an explanatory device to deal with cognitive phenomena.
- They take in particular closed world reasoning (CWR) to be a fruitful formal framework.
- They have applied CWR to a number of experimental results: Wason selection task, suppression task etc.
- But they have not looked into the ‘belief bias’ experiments. This is what I want to do today.

Introduction

- Stenning and van Lambalgen (2008) advocate the usefulness of non-monotonic logics as an explanatory device to deal with cognitive phenomena.
- They take in particular closed world reasoning (CWR) to be a fruitful formal framework.
- They have applied CWR to a number of experimental results: Wason selection task, suppression task etc.
- But they have not looked into the ‘belief bias’ experiments. This is what I want to do today.

Introduction

- Stenning and van Lambalgen (2008) advocate the usefulness of non-monotonic logics as an explanatory device to deal with cognitive phenomena.
- They take in particular closed world reasoning (CWR) to be a fruitful formal framework.
- They have applied CWR to a number of experimental results: Wason selection task, suppression task etc.
- But they have not looked into the ‘belief bias’ experiments. This is what I want to do today.

Introduction

- Stenning and van Lambalgen (2008) advocate the usefulness of non-monotonic logics as an explanatory device to deal with cognitive phenomena.
- They take in particular closed world reasoning (CWR) to be a fruitful formal framework.
- They have applied CWR to a number of experimental results: Wason selection task, suppression task etc.
- But they have not looked into the ‘belief bias’ experiments. This is what I want to do today.

Belief bias

- The tendency subjects have “to endorse arguments whose conclusions they believe and to reject arguments whose conclusions they disbelieve, irrespective of their actual validity”.
- The tendency to reason towards the confirmation of the beliefs we already hold.
- A ‘fundamental computational bias’ (Stanovich): “the tendency to automatically bring prior knowledge to bear when solving problems”.
- Conflict between ‘logic’ and ‘belief’.

Belief bias

- The tendency subjects have “to endorse arguments whose conclusions they believe and to reject arguments whose conclusions they disbelieve, irrespective of their actual validity”.
- The tendency to reason towards the confirmation of the beliefs we already hold.
- A ‘fundamental computational bias’ (Stanovich): “the tendency to automatically bring prior knowledge to bear when solving problems”.
- Conflict between ‘logic’ and ‘belief’.

Belief bias

- The tendency subjects have “to endorse arguments whose conclusions they believe and to reject arguments whose conclusions they disbelieve, irrespective of their actual validity”.
- The tendency to reason towards the confirmation of the beliefs we already hold.
- A ‘fundamental computational bias’ (Stanovich): “the tendency to automatically bring prior knowledge to bear when solving problems”.
- Conflict between ‘logic’ and ‘belief’.

Belief bias

- The tendency subjects have “to endorse arguments whose conclusions they believe and to reject arguments whose conclusions they disbelieve, irrespective of their actual validity”.
- The tendency to reason towards the confirmation of the beliefs we already hold.
- A ‘fundamental computational bias’ (Stanovich): “the tendency to automatically bring prior knowledge to bear when solving problems”.
- Conflict between ‘logic’ and ‘belief’.

Plan of the talk

- Present experimental data
- Present the notions of preferred model and preferential consequence
- Discuss the experimental data in light of these concepts

1. Experimental data

Experiment on belief-bias (Evans et al 1983)

Valid-believable	Valid-unbelievable	Invalid-believable	Invalid-unbelievable
No police dogs are vicious.	No nutritional things are inexpensive.	No addictive things are inexpensive.	No millionaires are hard workers.
Some highly trained dogs are vicious.	Some vitamin tablets are inexpensive.	Some cigarettes are inexpensive.	Some rich people are hard workers.
Therefore, some highly trained dogs are not police dogs.	Therefore, some vitamin tablets are not nutritional.	Therefore, some addictive things are not cigarettes.	Therefore, some millionaires are not rich people.

Results

Percentage of arguments accepted as valid:

	Believable conclusion	Unbelievable conclusion
Valid	89	56
Invalid	71	10

* Clearly, prior beliefs are typically activated when subjects are drawing inferences or evaluating (the correctness of) arguments.

Results

Percentage of arguments accepted as valid:

	Believable conclusion	Unbelievable conclusion
Valid	89	56
Invalid	71	10

* Clearly, prior beliefs are typically activated when subjects are drawing inferences or evaluating (the correctness of) arguments.

Syllogisms with familiar vs. unfamiliar content
(Sá, West & Stanovich 1999)

All living things need water.
Roses need water.
Thus, roses are living things.
=> 32% of logically ‘correct’ responses

All animals of the hudon class are ferocious.
Wampets are ferocious.
Thus, wampets are animals of the hudon class.
=> 78% of logically ‘correct’ responses

Syllogisms with familiar vs. unfamiliar content

(Sá, West & Stanovich 1999)

All living things need water.
Roses need water.
Thus, roses are living things.
=> 32% of logically ‘correct’ responses

All animals of the hudon class are ferocious.
Wampets are ferocious.
Thus, wampets are animals of the hudon class.
=> 78% of logically ‘correct’ responses

Syllogisms with familiar vs. unfamiliar content
(Sá, West & Stanovich 1999)

All living things need water.
Roses need water.
Thus, roses are living things.
=> 32% of logically ‘correct’ responses

All animals of the hudon class are ferocious.
Wampets are ferocious.
Thus, wampets are animals of the hudon class.
=> 78% of logically ‘correct’ responses

Syllogisms with familiar vs. unfamiliar content

(Sá, West & Stanovich 1999)

All living things need water.
Roses need water.
Thus, roses are living things.

=> 32% of logically ‘correct’ responses

All animals of the hudon class are ferocious.
Wampets are ferocious.
Thus, wampets are animals of the hudon class.

=> 78% of logically ‘correct’ responses

Conclusion production tasks (Oakhill & Johnson-Laird 1985)

Some of the actresses are not beautiful.
All of the women are beautiful.

Some of the A are not B
All of the C are B
Thus, some of the A are not C

Some of the actresses are not women (correct)	38%
No valid conclusion (error)	46%
Other errors	16%

Conclusion production tasks (Oakhill & Johnson-Laird 1985)

Some of the actresses are not beautiful.
All of the women are beautiful.

Some of the A are not B
All of the C are B
Thus, some of the A are not C

Some of the actresses are not women (correct)	38%
No valid conclusion (error)	46%
Other errors	16%

Conclusion production tasks (Oakhill & Johnson-Laird 1985)

Some of the women are not beautiful
All of the beautiful people are actresses

Some of the A are not B
All of the B are C
NO CONCLUSION

No valid conclusion (correct)	17%
Some of the women are not actresses (error)	46%
Other errors	37%

Conclusion production tasks (Oakhill & Johnson-Laird 1985)

Some of the women are not beautiful
All of the beautiful people are actresses

Some of the A are not B
All of the B are C
NO CONCLUSION

No valid conclusion (correct)	17%
Some of the women are not actresses (error)	46%
Other errors	37%

2. Preferred models and preferential consequence

Preferred models and preferential consequence

- (Shoham 1987) proposed a unifying framework for non-monotonic logics.
- It is general in that it can accommodate different preference criteria, thus generating different non-monotonic logics.
- Non-monotonic logics result from associating a standard logic with a preference relation on models.

Preferred models and preferential consequence

- (Shoham 1987) proposed a unifying framework for non-monotonic logics.

- It is general in that it can accommodate different preference criteria, thus generating different non-monotonic logics.

- Non-monotonic logics result from associating a standard logic with a preference relation on models.

Preferred models and preferential consequence

- (Shoham 1987) proposed a unifying framework for non-monotonic logics.

- It is general in that it can accommodate different preference criteria, thus generating different non-monotonic logics.

- Non-monotonic logics result from associating a standard logic with a preference relation on models.

Generating a non-monotonic logic

- Take a standard, monotonic logic $\mathscr{L}$: for all A, B and C in $\mathscr{L}$, if A => C, then also A $\wedge$ B => C
- Define a strict partial order $\angle$ on the models of $\mathscr{L}$: $M_1 \angle M_2$ means that M_2 is preferred over M_1.
- $\mathscr{L}_\angle$ is the non-monotonic logic generated from $\mathscr{L}$ and $\angle$.

Generating a non-monotonic logic

- Take a standard, monotonic logic $\mathscr{L}$: for all A, B and C in $\mathscr{L}$, if A => C, then also A $\wedge$ B => C

- Define a strict partial order $\angle$ on the models of $\mathscr{L}$: $M_1 \angle M_2$ means that M_2 is preferred over M_1.

- $\mathscr{L}_\angle$ is the non-monotonic logic generated from $\mathscr{L}$ and $\angle$.

Generating a non-monotonic logic

- Take a standard, monotonic logic $\mathscr{L}$: for all A, B and C in $\mathscr{L}$, if A => C, then also A $\wedge$ B => C

- Define a strict partial order $\angle$ on the models of $\mathscr{L}$: $M_1 \angle M_2$ means that M_2 is preferred over M_1.

- $\mathscr{L}_{\angle}$ is the non-monotonic logic generated from $\mathscr{L}$ and $\angle$.

Preferred models and preferential consequence

- A model M preferentially satisfies A ($M \models_\angle A$) if $M \models A$ and if there is no other model M’ such that $M \angle M'$ and $M' \models A$. M is a *preferred model* of A.
- A is a *preferential consequence* of B ($A \Rightarrow_\angle B$) if, for any M, if $M \models_\angle A$, then $M \models B$; that is, if the models of B (preferred or otherwise) are a superset of the preferred models of A.
- $\mathcal{L}_\angle$ is non-monotonic because $A \wedge B$ may have preferred models that are not preferred models of A (the two classes may be completely disjoint).

Preferred models and preferential consequence

- A model M preferentially satisfies A ($M \models_{\angle} A$) if $M \models A$ and if there is no other model M’ such that $M \angle M'$ and $M' \models A$. M is a *preferred model* of A.
- A is a *preferential consequence* of B ($A =>_{\angle} B$) if, for any M, if $M \models_{\angle} A$, then $M \models B$; that is, if the models of B (preferred or otherwise) are a superset of the preferred models of A.
- $\mathscr{L}_{\angle}$ is non-monotonic because $A \wedge B$ may have preferred models that are not preferred models of A (the two classes may be completely disjoint).

Preferred models and preferential consequence

- A model M preferentially satisfies A ($M \models_{\angle} A$) if $M \models A$ and if there is no other model M' such that $M \angle M'$ and $M' \models A$. M is a *preferred model* of A.
- A is a *preferential consequence* of B ($A =>_{\angle} B$) if, for any M, if $M \models_{\angle} A$, then $M \models B$; that is, if the models of B (preferred or otherwise) are a superset of the preferred models of A.
- $\mathscr{L}_{\angle}$ is non-monotonic because $A \wedge B$ may have preferred models that are not preferred models of A (the two classes may be completely disjoint).

Preferred models and belief bias

- What are the ‘preferred models’ of a human reasoner? The situations that accord with her prior beliefs and background knowledge about the world.
- The relation of preference is defined by the general state of prior beliefs.
- We can generalize the idea of a preferred model to the notion of *a class* of preferred models, so that the assumption of uniqueness is discarded.
- But even for classes of models, the assumption of a strict partial order of preference is an idealization.

Preferred models and belief bias

- What are the ‘preferred models’ of a human reasoner? The situations that accord with her prior beliefs and background knowledge about the world.
- The relation of preference is defined by the general state of prior beliefs.
- We can generalize the idea of a preferred model to the notion of *a class* of preferred models, so that the assumption of uniqueness is discarded.
- But even for classes of models, the assumption of a strict partial order of preference is an idealization.

Preferred models and belief bias

- What are the ‘preferred models’ of a human reasoner? The situations that accord with her prior beliefs and background knowledge about the world.
- The relation of preference is defined by the general state of prior beliefs.
- We can generalize the idea of a preferred model to the notion of *a class* of preferred models, so that the assumption of uniqueness is discarded.
- But even for classes of models, the assumption of a strict partial order of preference is an idealization.

Preferred models and belief bias

- What are the ‘preferred models’ of a human reasoner? The situations that accord with her prior beliefs and background knowledge about the world.
- The relation of preference is defined by the general state of prior beliefs.
- We can generalize the idea of a preferred model to the notion of *a class* of preferred models, so that the assumption of uniqueness is discarded.
- But even for classes of models, the assumption of a strict partial order of preference is an idealization.

3. Discussion

Two ‘unusual’ patterns

- Subjects draw inferences to ‘conclusions’ that do not follow deductively from the premises if they accord with prior belief.

- Subjects refuse to draw inferences to conclusions that do follow deductively from the premises if they go against prior belief.

Two ‘unusual’ patterns

- Subjects draw inferences to ‘conclusions’ that do not follow deductively from the premises if they accord with prior belief.

- Subjects refuse to draw inferences to conclusions that do follow deductively from the premises if they go against prior belief.

Inferences to ‘conclusions’ I

Some of the women are not beautiful: ψ
> All of the beautiful people are actresses: φ

- If a premise is not part of the prior state of belief, an update is required: $M \otimes \varphi = M^*$

- But in M* it is still the case that χ: ‘some of the women are not actresses (background information): $M^* \models \chi$

- So $M^* \models_{\angle} \psi, \varphi$ and $M^* \models \chi$, thus $\psi, \varphi \Rightarrow_{\angle} \chi$

Inferences to ‘conclusions’ I

Some of the women are not beautiful: ψ
> All of the beautiful people are actresses: φ

- If a premise is not part of the prior state of belief, an update is required: $M \otimes \varphi = M^*$

- But in M* it is still the case that χ: ‘some of the women’ are not actresses (background information): $M^* \models \chi$

- So $M^* \models_{\angle} \psi, \varphi$ and $M^* \models \chi$, thus $\psi, \varphi =>_{\angle} \chi$

Inferences to ‘conclusions’ I

Some of the women are not beautiful: ψ
> All of the beautiful people are actresses: φ

- If a premise is not part of the prior state of belief, an update is required: $M \otimes \varphi = M^*$

- But in M* it is still the case that χ: ‘some of the women are not actresses’ (background information): $M^* \models \chi$

- So $M^* \models_{\angle} \psi, \varphi$ and $M^* \models \chi$, thus $\psi, \varphi \Rightarrow_{\angle} \chi$

Inferences to ‘conclusions’ I

Some of the women are not beautiful: ψ
> All of the beautiful people are actresses: φ

- If a premise is not part of the prior state of belief, an update is required: $M \otimes \varphi = M^*$
- But in M* it is still the case that χ: ‘some of the women are not actresses’ (background information): $M^* \models \chi$
- So $M^* \models_{\angle} \psi, \varphi$ and $M^* \models \chi$, thus $\psi, \varphi \Rightarrow_{\angle} \chi$

Inferences to ‘conclusions’ II

All living things need water.
Roses need water.
Thus, roses are living things.

- This argument also satisfies the definition of preferential consequence (in all of the agent’s preferred models, roses are living things).
- Hypothesis: the addition of another premise, ‘some things that need water are not living things’ might make some subjects retract the conclusion.
- *Awareness* may be an important element.

Inferences to ‘conclusions’ II

All living things need water.
Roses need water.
Thus, roses are living things.

- This argument also satisfies the definition of preferential consequence (in all of the agent’s preferred models, roses are living things).
- Hypothesis: the addition of another premise, ‘some things that need water are not living things’ might make some subjects retract the conclusion.
- *Awareness* may be an important element.

Inferences to ‘conclusions’ II

All living things need water.
Roses need water.
Thus, roses are living things.

- This argument also satisfies the definition of preferential consequence (in all of the agent’s preferred models, roses are living things).
- Hypothesis: the addition of another premise, ‘some things that need water are not living things’ might make some subjects retract the conclusion.
- *Awareness* may be an important element.

Inferences to ‘conclusions’ III

All animals of the hudon class are ferocious.
Wampets are ferocious.
Thus, wampets are animals of the hudon class.

- The agent has no background knowledge about the hudon class or wampets: in her preferred models, the conclusion neither holds nor does not hold.
- So she cannot resort to preferential reasoning to judge the validity of this argument.
- Some other reasoning strategy is called upon, which explains the discrepancy in the results.

Inferences to ‘conclusions’ III

All animals of the hudon class are ferocious.
Wampets are ferocious.
Thus, wampets are animals of the hudon class.

- The agent has no background knowledge about the hudon class or wampets: in her preferred models, the conclusion neither holds nor does not hold.
- So she cannot resort to preferential reasoning to judge the validity of this argument.
- Some other reasoning strategy is called upon, which explains the discrepancy in the results.

Inferences to ‘conclusions’ III

All animals of the hudon class are ferocious.
Wampets are ferocious.
Thus, wampets are animals of the hudon class.

- The agent has no background knowledge about the hudon class or wampets: in her preferred models, the conclusion neither holds nor does not hold.
- So she cannot resort to preferential reasoning to judge the validity of this argument.
- Some other reasoning strategy is called upon, which explains the discrepancy in the results.

Refusing to draw inferences to conclusions

- Preferential reasoning is not able to explain why subjects refuse to validly draw a conclusion when it is unbelievable.
- After all, if A => B, then A =>$_{\angle}$ B, as the preferred models of A are also models of A *tout court*.
- Since the models of B form a superset of the models of A, they also form a superset of the preferred models of A.
- Hypotheses: the class of preferred models satisfying the premises is empty; it is inconsistent; there are no preferred models of the *conclusion*.

Refusing to draw inferences to conclusions

- Preferential reasoning is not able to explain why subjects refuse to validly draw a conclusion when it is unbelievable.
- After all, if A => B, then A =>$_{\angle}$ B, as the preferred models of A are also models of A *tout court*.
- Since the models of B form a superset of the models of A, they also form a superset of the preferred models of A.
- Hypotheses: the class of preferred models satisfying the premises is empty; it is inconsistent; there are no preferred models of the *conclusion*.

Refusing to draw inferences to conclusions

- Preferential reasoning is not able to explain why subjects refuse to validly draw a conclusion when it is unbelievable.
- After all, if A => B, then A =>$_{\angle}$ B, as the preferred models of A are also models of A *tout court*.
- Since the models of B form a superset of the models of A, they also form a superset of the preferred models of A.
- Hypotheses: the class of preferred models satisfying the premises is empty; it is inconsistent; there are no preferred models of the *conclusion*.

Refusing to draw inferences to conclusions

- Preferential reasoning is not able to explain why subjects refuse to validly draw a conclusion when it is unbelievable.
- After all, if A => B, then A =>$_{\angle}$ B, as all the preferred models of A are also models of A *tout court*.
- Since the models of B form a superset of the models of A, they also form a superset of the preferred models of A.
- Hypotheses: the class of preferred models satisfying the premises is empty; it is inconsistent; there are no preferred models of the *conclusion*.

Conclusions

- Non-monotonic logics provide a fruitful framework to think about the phenomenon of belief bias.
- The notion of preferred models is a natural conceptualization of the idea of bringing prior belief to bear, of ‘holding on’ to the beliefs we already have.
- But this approach only offers a partial explanation of the phenomena; it cannot explain why subjects refuse to draw unbelievable conclusions.
- Elements to be included: awareness of bits of information, the role of the preferred models of the conclusion.

Conclusions

- Non-monotonic logics provide a fruitful framework to think about the phenomenon of belief bias.
- The notion of preferred models is a natural conceptualization of the idea of bringing prior belief to bear, of ‘holding on’ to the beliefs we already have.
- But this approach only offers a partial explanation of the phenomena; it cannot explain why subjects refuse to draw unbelievable conclusions.
- Elements to be included: awareness of bits of information, the role of the preferred models of the conclusion.

Conclusions

- Non-monotonic logics provide a fruitful framework to think about the phenomenon of belief bias.
- The notion of preferred models is a natural conceptualization of the idea of bringing prior belief to bear, of ‘holding on’ to the beliefs we already have.
- But this approach only offers a partial explanation of the phenomena; it cannot explain why subjects refuse to draw unbelievable conclusions.
- Elements to be included: awareness of bits of information, the role of the preferred models of the conclusion.

Conclusions

- Non-monotonic logics provide a fruitful framework to think about the phenomenon of belief bias.
- The notion of preferred models is a natural conceptualization of the idea of bringing prior belief to bear, of ‘holding on’ to the beliefs we already have.
- But this approach only offers a partial explanation of the phenomena; it cannot explain why subjects refuse to draw unbelievable conclusions.
- Elements to be included: awareness of bits of information, the role of the preferred models of the conclusion.