Moore's paradox and hedging with 'I believe': An attempt.

Sven Lauer
University of Konstanz

Questioning Speech Acts
Konstanz, September 14 – 16, 2017

Wanted: A compositional analysis that jointly predicts two well-known observations:

- The fact that *I believe* often (but not always) **functions as a hedge**:

(1) I believe/think it is raining.
$\rightsquigarrow$ Sp is not certain that it is raining.

- **Moore's paradox**, i.e. the infelicity / contradictoriness of sentences like (2).

(2) #It is not raining but I believe it is (raining).

The desideratum of **compositionality** amounts to this:

- (3) should have the same kind of content as (4) and (5), modulo the belief subject / tense.

(3) I believe p.
(4) John believes p.
(5) I believed that p.

- All declarative sentences, including sentences of the form in (3) should receive (as declaratives) **a uniform sentential force**.
  - The different effects of asserting, e.g., (6) and (7) should result from their different contents.

(6) It is raining.
(7) I believe it is raining

Plot

- In **Section 1**, we take a closer look at the two phenomena and bring out a **tension** between (natural exaplanations of) them.
- In **Section 2**:
  - I take stock, formulate two more desiderata for an analysis of *(I) believe*-sentences.
  - Make plausible that the tensioin between the two phenomena call for a theory of **graded belief**.
  - Briefly say way I don't think **probability theory** is the way to go.
- In **Section 3**, I introduce Spohn's **ranking theory**.
- In **Section 4**, I use this theory in the interpretation of a **propositional language with belief-operators**.
- In **Section 5**, I review the **considerable success** with respect to the desiderata identified in the first half of the talk.
- In the **rest of the day**, various speakers (Klecha, Greenberg/Lavi, Mari) will surely challenge my analysis by discussing various facts that my analysis does not cover.

1 Two observations, and a dilemma.

1.1 Hedging with 'I believe'

- *I believe that p* often indicates that the speaker is not entirely certain that p, cf. (1).

(1) I believe/think it is raining. $\mathsf{Bel}_{Sp}(p)$
$\rightsquigarrow$ Sp is not certain that p.

- However, this is not always the case, in particular the inference can be suspended by using an adverb like *firmly*:[1]

(8) I firmly believe that it is raining.
$\not\rightsquigarrow$ Sp is uncertain whether it is raining.

- A natural way to account for this is as an **implicature**, derived (roughly) as follows:
  - The speaker of (1) could have asserted *It is raining*, which is shorter/less complex.
  - He must have had a reason to do so.
  - Maybe he did not want to commit to *It is raining* to be true, and chose to only commit to the claim that he **believes** it is raining.
  - One reason for this may be that he is not quite sure whether it is raining.

↪ *I believe p* is a way of avoiding (fully) committing to p.

[1] Question: Is such modification possible with *I think* and friends? If not, why not?

1.2 Moore's paradox

- Moore's paradox (Moore 1942, 1944): (9) sounds 'contradictory'.

(9) It is raining but I don't believe it (is raining). $p \wedge \neg \mathsf{Bel}_{Sp}(p)$
or: $p \wedge \mathsf{Bel}_{Sp}(\neg p)$

- And yet, (9) appears to have a perfectly consistent content, cf. (10a) and (10b).

(10) a. It is raining but John does not believe it (is raining).
b. It was raining but I did not believe it (was raining).

- Throughout most of this talk, I will focus on the sentence in (2), to avoid having to worry about neg-raising.

(2) #It is not raising, but I believe it (is raining). $\neg p \wedge \mathsf{Bel}_{Sp}(p)$

- A natural way to account for this is as follows:
  - With uttering $\neg p$, the speaker commits to taking p to be false.
  - At the same time, with uttering *I believe p*, she commits to taking p to be true.
  - Thus (2) gives rise to incompatible commitments.
  - Hence it is odd to assert it (even though it could be true).

↪ *I believe p* commits the speaker to p.

1.3 A dilemma

- There is a tension between the two 'natural explanantions' just sketched.
- Let A_{Sp} be an operator representing the consequences of assertion of a declarative.
  - 'doxastic commitment' / commitment to believe (Condoravdi and Lauer 2011, Lauer 2013)
  - 'assertoric commitment' (Krifka 2014)
  - 'truth commitment' (Searle 1969, Krifka 2015)
  - ...
- Note: Such commitment is in principle **independent** from belief.

(11) $\mathsf{Bel}_{Sp}(p) \not\Leftarrow\not\Rightarrow \mathsf{A}_{Sp}(p)$

The dilemma:
Should the following 'mixed introspection' principle be valid?

(12) MIXED INTROSPECTION: $\mathsf{A}_{Sp}(\mathsf{Bel}_{Sp}(p)) \rightarrow \mathsf{A}_{Sp}(p)$

- The 'natural explanation' for **hedging with 'believe'** says **NO!**
  - Or else, asserting '*I believe p*' is not a way to avoid asserting p.
- The 'natural explanation' for **Moore's paradox** says **YES!**
  - Or else, '*I believe p*' does not create a commitment that is incompatible with the one triggered by '$\neg p$'.

- Aside:MIXED INTROSPECTION is independent from both introspection for belief (13) and introspection for assertoric commitment (14)/

(13) INTROSPECTION FOR BELIEF: $\mathsf{Bel}_{Sp}(\mathsf{Bel}_{Sp}(p)) \rightarrow \mathsf{Bel}_{Sp}(p)$

(14) INTROSPECTION FOR ASSERTORIC COMMITMENT: $\mathsf{A}_{Sp}(\mathsf{A}_{Sp}(p)) \rightarrow \mathsf{A}_{Sp}(p)$

  - (13) is commonly assumed for (rational) belief, especially in logical approaches.
  - (14) is a crucial assumption in Condoravdi and Lauer (2011)'s account of explicit performatives.

2 Diagnosis: Weakness and strength

- Intuitively, MIXED INTROSPECTION (12) should fail because '*I believe p*' (in some contexts) induces a **weaker** commitment than '*p*'.
- At the same time, the commitment should **not be too weak**.
  - It must be strong enough to explain **Moore's paradox**.
  - And it arguably should be strong enough to predicts the following two observations:

(15) **Strength**: *I believe p* and *I believe* $\neg p$ are incompatible.

(16) **Closure**: A speaker who asserts *I believe p* and *I believe q* is also committed to *I believe* $p \wedge q$.

- **Strength**, in particular, requires that the commitment induced by '*I believe p*' is stronger than that by *Might p*, cf. (17).

(17) It might be raining, but it might also not be raining.

2.1 Graded belief

- So we need a 'medium-strong' commitment for *I believe*-sentences.
- This motivates employing a theory of **graded belief** that allows us to distinguish more levels than possibility and necessity.
- Let us try **probability theory** (cf., e.g. Swanson 2006, Lassiter 2011, 2017 on epistemic *must*).
- Set aside compositionality and assume:

  1. '*I believe that* p' commits the speaker to $P_{Sp}(p) > \beta$.
  2. 'p' commits the speaker to $P_{Sp}(p) > \alpha$.

  where $\alpha, \beta \geq 0.5$.

- Such a theory is set-up to do well on **Moore's paradox** and **Strength**.
  - Because a probability distribution can assign probability > 0.5 to at most one of p and $\neg p$.
- However, such a probabilistic-threshold theory can deliver on at most one of **Hedging** and **Closure**.
  - To account for **Hedging**, it must be that $\beta < \alpha \leq 1$.
  - But then $\beta < 1$, and hence **Closure** is not accounted for.

Wanted: A theory of graded belief (and assertion) that meets the following desiderata:

(18) **Hedging:**
$\mathsf{Bel}_{Sp}(p)$ induces a wekaer commitment than p.

(19) **Moore's paradox:**
$\neg p$ and $\mathsf{Bel}_{Sp}(p)$ induce incompatible commitments.

(20) **Strength:**
$\mathsf{Bel}_{Sp}(p)$ and $\mathsf{Bel}_{Sp}(\neg p)$ induce incompatible commitments.

(21) **Closure:**
$\mathsf{Bel}_{Sp}(p)$ and $\mathsf{Bel}_{Sp}(q)$ jointly commit the agent to $\mathsf{Bel}_{Sp}(p \wedge q)$

3 Ranking Theory

- **Ranking theory** (Spohn 1988, 1990, 2012) is an alternative theory of graded belief.
- One of its advertised features is that it predicts closure for belief.
- So let's have a closer look.

Definition 1 (Pointwise ranking functions, after Spohn 2012, p. 70)**.** *Given a set of worlds W, a* **complete pointwise (negative) ranking function** *is any function $\kappa : W \to (\mathbb{N} \cup \{\infty\})$ such that $\kappa^{-1}(0) \neq \emptyset$.*

- A complete pointwise ranking function simply assigns each world a natural number (or ∞).
- The only constraint is that **some** worlds must be assigned 0.
- This is a 'negative' ranking function because the intended interpretation is that it measures the 'disbelief' in worlds.
  - $\kappa(w) = 0$ indicates that w is one of the ''most expected' worlds according to the belief agent.
  - $\kappa(v) > \kappa(w)$ indicates taht w is 'more expected' by the belief agent than w.

Definition 2 (Lift)**.** *Given a complete pointwise ranking function κ, its* **lift** *($\kappa^{\uparrow}$) is that function $\wp(W) \to (\mathbb{N} \cup \{\infty\})$ such that*

1. $\kappa^{\uparrow}(\emptyset) = \infty$
2. *for any non-empty $A \subseteq W$: $\kappa^{\uparrow}(A) = \min\{\kappa(w) \mid w \in A\}$*

. Note: It is guaranteed that $\kappa^{\uparrow}(W) = 0$. $\kappa^{\uparrow}$ is a 'completely minimitive negative ranking function'.

- Such a negative ranking function for propositions modes **disbelief in propositions**.
- I.e., $\kappa^{\uparrow}$ tells us of each proposition how 'surprising' it would be for the agent.
- We could work with negative ranking functions throughout, but **positive ranking functions** are more intuitive.

Definition 3 (Positive lift, after Spohn 2012, p. 75)**.** *Given a complete pointwise ranking function κ, its* **positive lift** *(κ^+) is that function $\wp(W) \to (\mathbb{N} \cup \{\infty\})$ such that for all non-emtpy $A \subseteq W$:*

$$\kappa^+(A) = \kappa^\dagger(W - A)$$

- The positive lift of a ranking function gives us a measure of **belief** (rather than disbelief) for a proposition. In particular, the following hold:

  1. The contradictory proposition always has rank zero: $\kappa^+(\emptyset) = 0$
  2. The tautological proposition always has infinite rank: $\kappa^+(W) = \infty$
  3. For any $A, B \subseteq W$: $\kappa^+(A \cap B) = \min(\kappa^+(A), \kappa^+(B))$.

     More generally: $\mathcal{B} \subseteq \wp(W)$: $\kappa^+(\bigcap \mathcal{B}) = \min\{\kappa^+(B) \mid B \in \mathcal{B}\}$.

  Any function that satisfies these properties (and is defined for all $A \subseteq W$) is called a **completely minimitive positive ranking function on** W.

- A useful thing to keep in mind: For any A **at least one of** A **and** $W - A$ **must have rank zero**.

4 The object language: Syntax and semantics.

4.1 Syntax

For simplicity, we use a standard propositional language, enriched with a family of modal operators for belief:

Definition 4 (Language)**.** *Let P and A be disjoint sets (of proposition letters and agent names). Then $\mathcal{L}_{P,A}$ is the smallest set such that*

1. $P \subseteq \mathcal{L}_{P,A}$ *(proposition letters are formulas)*
2. *If $\phi \in \mathcal{L}_{P,A}$, then $\neg\phi \in \mathcal{L}_{P,A}$.* *(negation of formulas)*
3. *If $\phi, \psi \in \mathcal{L}_{P,A}$, then $(\phi \wedge \psi) \in \mathcal{L}_{P,A}$.* *(conjunction of formulas)*
4. *If $\phi \in \mathcal{L}_{P,A}$ and $a \in A$, then $(\mathsf{Bel}_a\phi) \in \mathcal{L}_{P,A}$.* *(belief formulas)*

Other connectives introduced as the usual abbreviations.

- Thus we have arbitrary Boolean combinations.

  (22) $p \wedge (\mathsf{Bel}_a q)$
  (23) $\neg p \wedge \neg(\mathsf{Bel}_{Sp} \neg q)$

  etc.

- Belief operators can nest, regardless of the agent involved:

  (24) $(\mathsf{Bel}_a(\mathsf{Bel}_b p))$
  (25) $(\mathsf{Bel}_a(\mathsf{Bel}_a p))$

4.2 Semantics

Models are standard possible-worlds one, with two additions:

Definition 5 (Models)**.** *A* **model** *for $\mathcal{L}_{P,A}$ is a qadruple $M = \langle W, I, K, \beta \rangle$, such that*

1. *W is a set of possible worlds,*
2. *$I : W \times P \to \{0, 1\}$ assigns each world a valuation for the proposition letters.*
3. *K a function that assigns to each agent-world pair complete pointwise ranking function.*
4. *$\beta \in \mathbb{N}$ is the threshold for belief ascriptions.*[2]

[2]Of course, in a more realistic system, we probably would let β be a contextual parameter.

With this, we can define a standard propositional semantics. The only interesting clause is 4:

Definition 6 (Denotation function). *Given a model* $M = \langle W, I, K, \beta \rangle$, *the* **denotation function** $[\![\cdot]\!]^M : \mathcal{L}_{P,A} \to \wp(W)$ *is as follows:*

1. $[\![p]\!]^M = \{w \in W \mid I(w,p) = 1\}$, *for all* $p \in P$.
2. $[\![\neg\phi]\!]^M = W - [\![\phi]\!]^M$.
3. $[\![\phi \wedge \psi]\!]^M = [\![\phi]\!]^M \cap [\![\psi]\!]^M$.
4. $[\![\mathsf{Bel}_a \phi]\!]^M = \{w \in W \mid K(a,w)^+(\phi) > \beta\}$

4.3 Introspection

To guarantee introspection, we define *admissibility* for ranking functions and models.

Definition 7 (Admissibility of ranking functions). *Given* $M = \langle W, I, K, \theta \rangle$, *a pointwise ranking function* κ *is* **admissible for** $a \in A$ *iff*

$$\forall v \in \kappa^{-1}(0) : K(v,a) = \kappa$$

That is, κ is admissible for a only if a's ranking function is the same as κ all worlds in receiving a negative rank 0 (The 'core' of the ranking function.)[3]

Definition 8 (Admissibility for models). *A model is admissible iff* $\forall w \in W, a \in A : K(w,a)$ *is admissible for* a.

- That is, a model is admissible iff K only assigns admissible ranking functions.
- Admissibility ensures introspection, the following sense:

Fact 9. *(Collapse) For any admissible model* $M : [\![\mathsf{Bel}_a(\mathsf{Bel}_a\phi)]\!]^M \supseteq [\![\mathsf{Bel}_a\phi]\!]^M$ *for all* a, ϕ.

Proof. Suppose $w \in [\![\mathsf{Bel}_a(\mathsf{Bel}_a\phi)]\!]$. Then $K(a,w)^+([\![\mathsf{Bel}_a\phi]\!]) \geq \beta$ and hence for all v such that $K(a,w)(v) \leq \beta : v \in [\![\mathsf{Bel}_a\phi]\!]$. Let v' be such that $K(w,a)(v') = 0$. As we have just seen, $v' \in [\![\mathsf{Bel}_a\phi]\!]$. Hence $K(a,v')^+([\![\phi]\!]) \geq \beta$. But, by admissibility: $K(a,v') = K(a,w)$. So $K(a,w)^+([\![\phi]\!]) \geq \beta$. But then, $w \in [\![\mathsf{Bel}_a(\phi)]\!]$.

□

[3] This notion of admissibility is actually too weak to deal with belief revision. But it will do for our (static) purposes here.

5 Declarative force

5.1 Commitment states

- Fixing a model M for a language $\mathcal{L}_{P,A}$, the commitments an agent has are represented as **constraints on ranking functions**:

(26) A **commitment state** C_a for $a \in A$ is partial truth-function of pointwise ranking functions such that $C_a(\kappa)$ is defined iff κ is admissible for a.

- There are two distinguished commitment states:

(27) $\bot = \lambda\kappa.0$ (the contradictory state)

(28) $\top = \lambda\kappa.1$ (the uncommitted state)

5.2 Updates for commitment states

- Declarative force is modeled via the following update operation on commitment states:

(29) $C + \phi = \lambda\kappa.C(\kappa) \,\&\, \kappa^{+}(\llbracket\phi\rrbracket^{M}) > \alpha$ (declarative update)

- Support is standardly (Veltman 1996-style) defined as vacuous update:

(30) $C \vDash \phi$ iff $C + \phi = C$ (support)

6 Success

6.1 Hedging explained

- In general:

(31) $C_a + \mathsf{Bel}_a(\phi) \nvDash \phi$

- I.e., updating with $\mathsf{Bel}_a(\phi)$ does not commit the speaker to ϕ.
- This is so because the update with $\mathsf{Bel}_a(\phi)$ only requires (by admissibility) that

$$\kappa(\llbracket\phi\rrbracket) > \beta$$

- This does not exclude that $\kappa(\llbracket\phi\rrbracket) \leq \alpha$, since $\alpha > \beta$.

6.2 Moore's paradox explained

- For any agent a and commitment state C_a:

(32) $C_a + (\neg\phi \wedge \mathsf{Bel}_a\phi) = \bot$
↪ *It is not raining, but I believe it is* is inconsistent.

(33) $C_a + (\phi \wedge \mathsf{Bel}_a\neg\phi) = \bot$
↪ *It is raining, but I believe it is not raining.* is inconsistent.

(34) $C_a + (\phi \wedge \neg\mathsf{Bel}_a\phi) = \bot$
↪ *It is raining, but I don't believe it is raining.* is inconsistent.

- For (32) this is so because the first conjunct requires $\kappa^+(\phi) = 0$, but the second requires $\kappa^+ > \theta \geq 0$.
- Reasoning for the other cases is analogous.
- N.B.: It can easily be that $[\![\neg\phi \wedge (\mathsf{Bel}_a\phi)]\!] \neq \emptyset$.

7 Strength explained

- For any a and commitment state C_a:

(35) $C_a + (\mathsf{Bel}_a\phi) + (\mathsf{Bel}_a\neg\phi) = \bot$

- The first update requires $\kappa^+(\phi) > \beta \geq 0$.
- The second update requires $\kappa^+(\neg\phi) > \beta \geq 0$.
- But a ranking function can assign positive rank to at most one of ϕ and $\neg\phi$.

8 Closure explained

- For any a and commitmentstate C_a:

(36) $C + (\mathsf{Bel}_a\phi) + (\mathsf{Bel}_a\psi) \vDash \mathsf{Bel}_a(\phi \wedge \psi)$

- $\mathsf{Bel}_a\phi$ requires that $\kappa^+(\phi) > \beta$.
- $\mathsf{Bel}_a\psi$ requires that $\kappa^+(\psi) > \beta$.
- But then, it must also be that $\min(\kappa^+(\phi), \kappa(\psi)^+) > \beta$.
- $\mathsf{Bel}_a(\phi \wedge \psi)$ requires that
- But that is already required by $C_a + (\mathsf{Bel}_a\phi) + (\mathsf{Bel}_a\psi)$.

9 Conclusion

In summary:

- If we want to maintain what I have called the 'natural explanation' of **Moore's paradox** and **Hedging with 'believe'**, we need to employ a theory of **graded belief** to avoid the dilemma from Section 1.
- If we also want to account for **Closure**, then **probability theory will not do**.
- However, **ranking theory** gives us an elegant tool for accounting for all three facts (and **Strength**) at the same time.

Some questions:

- We'll hear (much) more about belief ascriptions later today and throgout this workshop (e.g. in Klecha and Mari's talks).
  - Can their observations be accounted for in a ranking-theoretic framework?
  - Or do their observation point to crucial weaknesses in that framework?
- I have talked about (categorical) commitment to **graded belief**.
  - Could we also do with **graded commitment** à la Greenberg/Lavi?
  - And could we do so **compositionally**?
  - (This would seem to require a 'speech-acty' analysis of *believe*?)
  - (They might like that. So might Krifka.)

References

Condoravdi, C. and Lauer, S.: 2011, Performative verbs and performative acts, *in* I. Reich, E. Horch and D. Pauly (eds), *Sinn and Bedeutung 15*, Universaar – Saarland University Press, Saarbrücken, pp. 149–164.

Krifka, M.: 2014, Embedding illocutionary acts, *in* T. Roeper and P. Speas (eds), *Recursion, Complexity in Cognition*, Vol. 43 of *Studies in Theoretical Psycholinguistics*, Springer, Berlin, pp. 125–155.

Krifka, M.: 2015, Bias in commitment space semantics: Declarative questions, negated questions, and question tags, *Semantics and Linguistic Theory* **25**, 328–345.

Lassiter, D.: 2011, *Measurement and Modality: The Scalar Basis of Modal Semantics*, PhD thesis, New York University.

Lassiter, D.: 2017, *Graded Modality: Qualitative and Quantitative Perspectives*, Oxford University Press.

Lauer, S.: 2013, *Towards a dynamic pragmatics*, PhD thesis, Stanford University.

Searle, J. R.: 1969, *Speech Acts: An essay in the philosophy of language*, Cambridge University Press, Cambridge, UK.

Spohn, W.: 1988, Ordinal conditional functions: A dynamic theory of epistemic states, *in* W. Harper and B. Skyrms (eds), *Causation in Decision, Belief Change, and Statistics (Volume II)*, Kluwer, Dordrecht, pp. 105–134.

Spohn, W.: 1990, A general non-probabilistic theory of inductive reasoning, *in* R. Shachter, T. Levitt, J. Lemmer and L. Kanal (eds), *Uncertainty in Artificial Intelligence (Volume 4)*, Association for Uncertainty in Artificial Intelligence, Amsterdam, pp. 149–158.

Spohn, W.: 2012, *The laws of belief*, Oxford University Press, Oxford, UK.

Swanson, E.: 2006, *Interactions with context*, PhD thesis, Massachussetts Institute of Technology.

Veltman, F.: 1996, Defaults in update semantics, *Journal of Philosophical Logic* **25**, 221–261.