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

Sven Lauer, University of Konstanz

Wanted: A compositional analysis that jointly predicts two well-known observations: (i) *'I believe'* is frequently used as a **hedge**, cf. (1) (ii) **Moore's paradox**, cf. (2). The desideratum of compositionality amounts to this: *'I believe p'* should have the same kind of content as *'John believes p'*, modulo the belief subject; and *'p'* and *'I believe p'* should be assigned the same conventional force, with their different effects following form their different contents.

Hedging with 'I believe'. *'I believe that p'* often indicates that the speaker is not entirely certain that p, cf. (1). A natural way to account for this as an **implicature**, derived on the basis of the fact that the speaker chose to say *'I believe that p'*, instead of *'p'*. One reason for doing so may be that she does not want to commit to p, and instead only commmits to the claim that she believes that p.

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

Moore's paradox (Moore 1942, 1944). (2) seems to have a perfectly consitent content (cf. *'It is not raining, but John believes it is'*), yet it sounds "incoherent" (We focus on (2) instead of the (classic) *'It is raining but I don't believe it'* to side-step questions about neg-raising). A natural way to account for this is the following: With uttering $\neg p$, the speaker commits to taking p to be false, but with uttering $\mathsf{Bel}_{Sp}(p)$, she commits to taking p to be true. Thus, (2) induces incompatible commitments.

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

A dilemma. There is a tension between the two 'natural explanations' sketched above. To bring this tension out more clearly, let A_{Sp} be an operator representing the consequences of assertion of a declarative ('doxastic commitment', Condoravdi and Lauer 2011, Lauer 2013, 'assertoric commitment' Krifka 2014, 'truth commitment' Searle 1969, Krifka 2015, etc.). Then we are faced with the question whether the 'mixed introspection' principle in (3) should be valid (N.B., mixed introspection is in principle independent from introspection for 'believe': $\mathsf{Bel}_{Sp}(\mathsf{Bel}_{Sp}(p)) \rightarrow \mathsf{Bel}_{Sp}(p)$).

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

The 'natural explanation' for **hedging with 'believe'** requires that (3) be **not valid** (else, asserting 'I believe p' is not a way to avoid asserting p). On the other hand, the 'natural explanation' for **Moore's paradox** seems to rely on the assumption that (3) **is valid**.

Diagnosis. Intuitively, (3) should fail because asserting 'I believe that p' induces (at least in some contexts) a **weaker** commitment than asserting 'p'. At the same time, this commitment should **not be too weak**: It must be strong enough to explain Moore's paradox, and also (arguably) the two observations labelled **Strength** and **Closure** below. In particular, **Strength** requires that the commitment induced by 'I believe p' is stronger than that induced by *'Might p'*.

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

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

Graded belief. The need for 'medium-strong' commitment for *'believe'*-ascriptions motivates moving to a theory of graded belief, such as probability theory (cf. Swanson 2006, Lassiter 2011, 2017 on epistemic *'must'*). Setting aside compositionality, assume that *'I believe that p'* commits its speaker to her subjective probability distribution satisfying $P_{Sp}(p) > \theta$ and that asserting 'p' commits her to $P_{Sp}(p) > \alpha$, where $\alpha, \theta \geq 0.5$. Such a theory is set-up to do well on **Moore's paradox** and **Strength**, but it can only predict at most one of **Hedging** and **Closure**: To predict **Hedging**, it must be that $\theta < \alpha \leq 1$. But then, **Closure** does not hold.

We hence explore an account in terms of a different theory of graded belief: The **ranking theory** of Spohn (1988, 1990, 2012).

Definition 1 (Language)**.** *For P and A disjoint sets (proposition letters, agents), $\mathcal{L}_{P,A}$ is the smallest superset of P closed under negation $\neg$ and conjunction $\wedge$, s.t. if $\phi \in \mathcal{L}_{P,A}, a \in A$, then $(\mathsf{Bel}_a\phi) \in \mathcal{L}_{P,A}$.*

Definition 2 (Models)**.** *A* **model** *for $\mathcal{L}_{P,A}$ is a qadruple $M = \langle W, I, K, \theta\rangle$, such that W is a set of possible worlds, $I : W \times P \to \{0,1\}$ an interpretation function for the proposition letters, K a function that assigns to each agent-world pair complete pointwise ranking function and $\theta \in \mathbb{N}$.*

Definition 3 (Ranking functions, after Spohn 2012, p. 70/75)**.** *Given $M = \langle W, I, K, \theta\rangle$, a* **complete pointwise ranking function** *is a function $\kappa : W \to (\mathbb{N} \cup \{\infty\})$ such that $\kappa^{-1}(0) \neq \emptyset$. Given such a function κ, its* **positive lift** *(κ^+) is that function $\wp(W) \to (\mathbb{N} \cup \{\infty\})$ such that $\kappa^+(\emptyset) = 0$, $\kappa^+(W) = \infty$ and for any non-empty $A \subset W$: $\kappa^+(A) = \min\{\kappa(v) \mid v \in (W - A)\}$. For any κ, κ^+ is a completely minimitive ranking function. In particular, $\kappa^+(\bigcap \mathcal{B}) = \min\{\kappa^+(B) \mid B \in \mathcal{B}\}$ for all $\mathcal{B} \in \wp(W)$.*

Definition 4 (Denotation)**.** *Given a model $M = \langle W, I, K, \theta\rangle$, the* **denotation function** *$[\![\cdot]\!]^M : \mathcal{L}_{P,A} \to \wp(W)$ is as follows: Proposition letters are interpreted via I, Boolean combinations are interpreted in the usual way ($\neg$ as complement on W, $\wedge$ as $\cap$), and $[\![\mathsf{Bel}_a\phi]\!]^M = \{w \in W \mid K(a,w)^+(\phi) > \theta\}$.*

We define *admissibility* for ranking functions and models, ensuring introspection for belief (Fact 6).

Definition 5. *Given $M = \langle W, I, K, \theta\rangle$, a pointwise ranking function κ is* **admissible** *for $a \in A$ iff $\forall v \in \kappa^{-1}(0) : K(v,a) = \kappa$. A model is admissible iff $\forall w \in W, a \in A : K(w,a)$ is admissible for a.*

Fact 6. *(Collapse) For any admissible model M : $[\![\mathsf{Bel}_a(\mathsf{Bel}_a\phi)]\!]^M \supseteq [\![\mathsf{Bel}_a\phi]\!]^M$ for all a, ϕ.*

Declarative force Fixing a model M for a language $\mathcal{L}_{P,A}$, the commitments an agent has are represented as *constraints on ranking functions*. Thus a **commitment state** 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 $\bot = \lambda\kappa.0$ (the contradictory state) and $\top = \lambda\kappa.1$ (the uncommitted state). We further define the **update to commitment states** that happens when a speaker utters a declarative sentence as in (4), and, in terms of it, a notion of **support** à la Veltman (1996) in (5).

(4) $\quad C + \phi = \lambda\kappa.C(\kappa)\ \&\ \kappa^+(\phi) > \alpha$ $\qquad$ (5) $\quad C \vDash \phi$ iff $C + \phi = C$

Success of the account is witnessed by the following three facts:

Fact 7 (Hedging with *'belief'* explained)**.** *For any C_a: $C_a + \mathsf{Bel}_a(\phi) \nvDash \phi$ unless $C_a \vDash \phi$.*

This is so because $C_a + (\mathsf{Bel}_a\phi)$ only requires (in virtue of admissibility) that $\kappa^+(\phi) > \theta$, while ϕ requires $\kappa^+(\phi) > \alpha$, and $\theta < \alpha$.

Fact 8 (Moore's paradox explained)**.** *For any agent a and commitment state C_a :*

(i) $C_a + (\neg\phi \wedge \mathsf{Bel}_a\phi) = \bot$ $\qquad$ *(ii)* $C_a + (\phi \wedge \mathsf{Bel}_a\neg\phi) = \bot$ $\qquad$ *(iii)* $C_a + (\phi \wedge \neg\mathsf{Bel}_a\phi) = \bot$

For (i) 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 $[\![\neg\phi \wedge (\mathsf{Bel}_a\phi)]\!] \neq \emptyset$.

Fact 9 (Strength and closure explained)**.** *For any agent a and commitment state C_a:*

(i) $C_a + (\mathsf{Bel}_a\phi) + (\mathsf{Bel}_a\neg\phi) = \bot$ $\qquad$ *(ii)* $C + (\mathsf{Bel}_a\phi) + (\mathsf{Bel}_a\psi) \vDash \mathsf{Bel}_a(\phi \wedge \psi)$

For (i): The first update requires $\kappa^+(\phi) > \theta \geq 0$, the second $\kappa^+(\neg\phi) > \theta \geq 0$. But a ranking function can assign positive rank to at most one of ϕ and $\neg\phi$. For (ii): $\mathsf{Bel}_a(\phi \wedge \psi)$ requires that $\min(\kappa^+(\phi), \kappa(\psi)^+) > \theta$, but that is already required by $C_a + (\mathsf{Bel}_a\phi) + (\mathsf{Bel}_a\psi)$.

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 quetions, 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.