I'm Mistaken

Jeremy Goodman (Johns Hopkins University)

National University of Singapore – October 12, 2023

1 The puzzle of modesty

The following are inconsistent (see appendix A):

CON Your beliefs are consistent.
$\Diamond\forall p(@Bp \to p)$

MOD You believe that something you believe is false.
$B\exists p(Bp \wedge \neg p)$

NI You are negatively introspective.
$\forall p(\neg Bp \to B\neg Bp)$

Proof: If you're consistent, then in some world w everything you actually believe is true. If you're modest, then you actually believe that you have a false belief. So in w you do have a false belief – call it p. You don't actually believe p, since then p would have to be true in w, which it isn't. And you don't actually believe that you don't believe p, since then in w you would have to not believe p, but you do. So p it is a counterexample to your being negatively introspective.

This is very puzzling!
People like us aren't perfectly consistent or perfectly introspective. But idealized agents who were could still make mistakes through misperception, misinformation, and misleading evidence, and know this about themselves. Shouldn't they be modest?

2 Solution: reject negative introspection?

- We can only believe propositions that we understand(/comprehend). And for many propositions p, we neither understand p nor understand the proposition that we don't believe p.
- Perhaps surprisingly: making allowance for such failures of negative introspection suffices to reconcile logical and introspective perfection (as in appendix B) with modesty; see appendix C.

The carpenter (Peter Fritz, p.c.)
A carpenter cuts boards into planks all day. They have superhuman logical and introspective powers, but they are just as susceptible to misperception, misremembering, and being misinformed as anyone else. Knowing this, they believe that for some n, x, y, they falsely believe that they cut plank n to $x \times y$ cm.

Problem: The carpenter understands/comprehends all propositions of the form *I cut plank n to $x \times y$ cm*, and it is non-contingent which propositions these are. So their being logically and introspectively perfect in the sense of the theory in appendix B is inconsistent with their specific kind of modesty. See appendix D.

3 Are preface writers inconsistent?

The preface (Makinson, 1965)
You believe every claim in your book, including the claim (in the preface) that the book contains at least one error.

Let $p_1, \ldots, p_n$ be the claims in the book. It is possible that $p_1, \ldots, p_n$ all be true (and in the book) provided a new false claim q is written in the book too. So you needn't be inconsistent. (Evnine, 1999)

- Your beliefs would be inconsistent if you believed that every claim in the book was either p_1 or $\ldots$ or p_n. But this claim
  1. doesn't follow from what we've assumed you believe; and
  2. is too complicated for an ordinary person to understand.

4 Fragmented inconsistency

Reunion
You greet everyone at the door of your school reunion. A sensor keeps a count of the attendees. Afterwards, you remember everyone who came, and any two people who came you know are different people. You check the sensor to see how many people came. Unbeknownst to you, a freak malfunction caused it to undercount by one person.

Test Scores
You like to waste time by taking tests on random trivia. The way the tests work is that you skip questions you are unsure about, and after an hour the test ends and you get your score. You typically get around 97% correct. This time, you get 99/100. Every question you are asked you remember having answered, and you still believe all of your answers (after all, you did better than expected).

- In both **Reunion** and **Test Scores** you are inconsistent, and the inconsistency is spread over a large number of beliefs, none of which has book-level complexity.

No deep link between modesty and inconsistency

Test Scores might seem to suggest a connection here, but:

1. in **Reunion** you are inconsistent in the same way without necessarily being modest;
2. the same is true in a variant of **Test Scores** where you are misinformed of your score, and told you got 99/99; and
3. in the other direction, in a variant of **Test Scores** where you are misinformed of your score, and told you got 100/101, you will be modest but needn't be inconsistent.

How to be an inconsistent preface author

- Count the number of claims in your book.
- Memorize the book, so that you know it starts with p_1, that it ends with p_n, and that p_i is followed by p_{i+1} (for $1 \leq i < n$).

5 Modesty and the nature of belief

5.1 Hintikka semantics & euclidean accessibility

Many models of rational belief have the following abstract structure:

1. S believes that p in w if and only if p is true in all worlds that are doxastically accessible from w for S (Hintikka, 1962); and
2. there is a factor F (one's internal state, one's evidence, etc.) such that
   (a) if v is doxastically accessible from w for S, then w and v agree about how S is in respect F, and
   (b) if w and v agree about how S is in respect F, then they agree on which worlds are doxastically accessible for S.

Example: S believes that p in w if and only if p is true in the *most normal* worlds that agree with w about S's internal state (Stalnaker, 1984) or about S's evidence (Goodman and Salow, 2023).

Problem: (2) entails that doxastic accessibility is *euclidean* (if v and u are both accessible from w, then they are accessible from each other); given (1), this entails that agents are negatively introspective.

5.2 The Representational Theory of Thought

If you believe that you have a false belief, then you do have a false belief (Prior, 1961). This generates a version of the liar paradox.

Self-Fulfilling Modesty
Alice has only true beliefs and has never considered whether she has any false beliefs. Then she reflects on human fallibility and comes to believe that she has a false belief, without otherwise changing her opinions.

This is very puzzling!
All of Alice's old beliefs are true, and so is her new belief that she has a false belief. So in becoming modest she must have formed another different, false belief. What it is? To sharpen the puzzle, consider:

RTT: S believes that p if and only if some mental representation r both expresses p and is in S's belief box.

RTT is inconsistent with the intuitively possible situation that, for some mental representation β (i.e., 'I believe something false'):

1. β expresses (in S's mind) that S believes something false;
2. β expresses (in S's mind) nothing else;
3. β is in S's belief box; and
4. all other representations in S's belief box expresses only truths.

6 A tentative solution

- Idea: the expressing relation between mental representations and propositions is both one-many and modally plastic (Dorr and Hawthorne, 2014, Dorr, 2020); see appendix E for a model.
- Mentalese 'believe' expresses infinitely many attitudes at once. The more attitudes it expresses, the more you understand, so the more you know you understand, so the stronger your evidence. This non-transparency of evidence suggests a solution to the problem of euclidean accessibility (see Remark 1 below).

A The puzzle of modesty formalized

We make the following assumptions about the interaction of $\Box$, @, propositional quantifiers and Boolean connectives:

RIG $\varphi \leftrightarrow \Box @\varphi$

$@_\forall$ $\Box(@\forall p\varphi \leftrightarrow \forall p@\varphi)$

$@_\rightarrow$ $\Box\forall p(@(\varphi \rightarrow \psi) \leftrightarrow (@\varphi \rightarrow @\psi))$

$@_\neg$ $\Box\forall p(@\neg\varphi \leftrightarrow \neg@\varphi)$

Recall that modesty, consistency and negative introspection are:

MOD $B\exists p(Bp \wedge \neg p)$

CON $\Diamond\forall p(@Bp \rightarrow p)$

NI $\forall p(\neg Bp \rightarrow B\neg Bp)$

Combining NI with RIG, $@_\forall$, $@_\rightarrow$ and $@_\neg$, in that order, we derive:

$$\Box\forall p(\neg@Bp \rightarrow @B\neg Bp)$$

Combined with CON, this implies that everything you actually believe being true is not merely possible, as CON states, but also compossible with everything you believe being something you actually believe:

$$\Diamond(\forall p(@Bp \rightarrow p) \wedge \forall p(Bp \rightarrow @Bp))$$

And given MOD and RIG, this is in turn compossible with your actual modesty:

$$\Diamond(\forall p(@Bp \rightarrow p) \wedge \forall p(Bp \rightarrow @Bp) \wedge @B\exists p(Bp \wedge \neg p))$$

But this is an impossibility: the first two conjuncts imply that everything you believe is true, while the first and third conjuncts imply that something you believe is false. One cannot be modest, consistent, and negatively introspective.

(Note that this result doesn't turn on puzzles about belief and actuality – e.g., that we sometimes believe the impossible under the guise of $\ulcorner @\varphi \urcorner$ when φ expresses a contingent falsehood – since @ is never embedded under B in the above derivation.)

B A theory of belief and understanding

PL every propositional tautology

K $B(\varphi \rightarrow \psi) \rightarrow (B\varphi \rightarrow B\psi)$

D $B\varphi \rightarrow \neg B\neg\varphi$

4 $B\varphi \rightarrow BB\varphi$

C4 $BB\varphi \rightarrow B\varphi$

5U $U\varphi \rightarrow (\neg B\varphi \rightarrow B\neg B\varphi)$ $\Leftarrow$ **this is the key thing**

Dist $\forall p(\varphi \rightarrow \psi) \rightarrow (\forall p\varphi \rightarrow \forall p\psi)$

Vac $\varphi \rightarrow \forall p\varphi$ where p is not free in φ

UI $\forall p\varphi \rightarrow \varphi[\psi/p]$ where ψ is free for p in φ

U $B\varphi \rightarrow U\varphi$

UB $U\varphi \rightarrow BU\varphi$

CUB $BU\varphi \rightarrow U\varphi$

CL $(U\varphi_1 \wedge \cdots \wedge U\varphi_n) \rightarrow U\psi$, where ψ is built from $\varphi_1, \ldots, \varphi_n$ using only $\neg, \wedge, B, U$ and quantifiers

MP If $\vdash \varphi \rightarrow \psi$ and $\vdash \varphi$, then $\vdash \psi$.

RN If $\vdash \varphi$, then $\vdash B\varphi$.

Gen If $\vdash \varphi$, then $\vdash \forall p\varphi$.

RE If $\vdash \varphi \leftrightarrow \psi$, then $\vdash \Phi \rightarrow \Phi[\psi/\varphi]$.

C The simple model

The following model establishes that the above theory is consistent with MOD and CON (assuming a standard interpretation of $\Diamond$ and @). Notice that doxastic accessibility is anti-euclidean: if $y, z \in R(x)$, then $z \notin R(y)$ or $z \notin R(y)$.

$W = \mathbb{R}$

$R(x) = \{y : y > x\}$

$[\![p]\!]^g = g(p)$

$[\![\neg]\!]^g(X) = W \backslash X$

$[\![\wedge]\!]^g(X)(Y) = X \cap Y$

$[\![U]\!]^g(X) = \{x \in W : y \in X \leftrightarrow z \in X \text{ for all } y, z \in R(x)\}$

$[\![B]\!]^g(X) = \{x \in W : R(x) \subseteq X\}$

$[\![\Diamond]\!]^g(X) = \{x \in W : X \neq \emptyset\}$

$[\![@]\!]^g(X) = \{x \in W : 0 \in X\}$

$[\![\forall p \varphi]\!]^g = \bigcap\{[\![\varphi]\!]^{g[p \to X]} : X \subseteq W\}$

Remark 1. In Goodman and Salow (2021, appendix A) we give a question-sensitive analysis of doxastic accessibility in terms of evidential probability. By allowing the question to be *world-relative*, we can derive the above doxastic accessibility relation, by letting $\mathcal{Q}_x = \{\{y : y \leq x\}, \{y : y > x\}\}$, and having *evidential accessibility* be $R_E(x) = \{y : y \geq x\}$ and requiring the prior probabilities to satisfy the condition that $Pr([x, y]) > 0$ iff $x \neq y$.

D The strengthened puzzle of modesty

MOD_X $B\exists p(Xp \wedge Bp \wedge \neg p)$

CON $\Diamond\forall p(@Bp \to p)$

5U $U\varphi \to (\neg B\varphi \to B\neg B\varphi)$

XU $\forall p(Xp \to Up)$

X$\Box$ $\Box\forall p(Xp \to @Xp)$

$Xp := p$ is a proposition of the form *I cut plank n to x × y cm*

E The mentalese model

Step 1: replace CL with the following weaker principle:

CL^- $U\varphi \to U\psi$, where ψ is built from φ using only $\neg, \wedge, B, U$ and quantifiers

Step 2: modify the model as follows:

$\pi_x = \{\bigcup X : X \subseteq \{\{y : y < x\}, \{x\}, \{y : y > x\}\}\}$

$[\![U]\!]^g(X) = \{x \in W : X \in \pi_y \text{ for some } y \leq x\}$

$[\![B]\!]^g(X) = [\![U]\!]^g(X) \cap \{x \in W : R(x) \subseteq X\}$

Step 3: consider the infinite family of interpretations $[\![\cdot]\!]_y$ defined like $[\![\cdot]\!]$ except for a different accessibility relation in the clause for B:

$R_y(x) = R(x)$ if $y < x$ and $= R(x) \cup \{x\}$ for $y \geq x$

$[\![B]\!]^g_y(X) = [\![U]\!]^g(X) \cap \{x \in W : R_y(x) \subseteq X\}$

Let T be the closed theorems of the theory of understanding and belief (with CL^- in place of CL) plus MOD. The idea is that this gives the sentences of mentalese in your belief box. Moreover, mentalese is semantically plastic and plenitudinous: in addition to having the interpretation assigned by $[\![\cdot]\!]$ in all worlds, it also has the interpretation assigned by $[\![\cdot]\!]_x$ in all worlds $y \geq x$.

Definition 2. You *understand* Y in x iff $x \in [\![U]\!](Y)$; you *believe* Y in x iff $x \in [\![B]\!](Y)$.

Definition 3. φ *expresses* Y in x iff φ is a closed sentence and either $Y = [\![\varphi]\!]$ or $Y = [\![\varphi]\!]_y$ for some $y \leq x$.

Proposition 4. You understand Y in x iff some φ expresses Y in x; you believe Y in x iff some $\varphi \in T$ expresses Y in x.