Belief in Belief and the Beetle-Box Metric*

James Somers

1. Introduction

As Daniel Dennett points out in *Breaking the Spell: Religion as a Natural Phenomenon*, believing that "democracy is good" is different from believing that "*belief in democracy* is good." Someone who held the former might write a pamphlet espousing the benefits of a democratic society, whereas someone who held the latter might see to the *distribution* of that pamphlet: theirs is a "second-order belief" or, as Dennett puts it, a "belief in belief."

Historically, religious beliefs tended to be of the first order: "you sacrifice an ox if you want it to rain" because "you really believe that the rain god won't provide rain unless you sacrifice an ox" (Dennett 227). Belief for its own sake appears to be a relatively recent invention. One explanation is that "the meme for faith exhibits *frequency-dependent fitness*: it flourishes particularly in the company of rationalistic memes"; since "rationalistic memes" have proliferated in recent centuries, so have calls for "blind faith" and belief in belief (231).

But we know that a person may believe in *belief in X* with or without believing *X* itself. If I were the editor of a left-wing journal, for instance, I might believe in left-wing principles and *also* in belief in those principles, because the latter is likely to improve the odds that my agenda is voted for. If, contrarily, I were the *owner* of that journal, I might still believe in belief in left-wing principles—because that sells copies—*without* believing in the principles themselves (because I am greedy capitalist). Which suggests the following quandary: how are we to tell who *actually* believes in God from those who merely believe in belief in God?

Dennett explores a naïve method—asking subjects "Does God exist?"—and explains why it is inadequate, but he does not pursue the question further. The goal of this paper is to continue where he left off.

My first task will be to articulate the problem—that one's true religious beliefs are not publicly available and, in some cases, not even articulable (Section 2). Dennett briefly likens this feature to Wittgenstein's "beetle in the box," an excellent analogy that I will pursue in much greater detail; in particular, I will introduce the *beetle-box metric* for classifying concepts (Section 3)—I argue that distinguishing belief in God (**B**) from belief in belief in God (**BB**) is easier for some God concepts than for others. To demonstrate the idea, I will divide the "beetle-box scale for *X*" into four broad classes:

1. Everyone has roughly the same concept *X* in their box, and *X* is (2. is not) articulable.
3. Everyone has a different concept *X* in their box, and *X* is (4. is not) articulable.

and for each of these explore different (progressively more difficult) "tests" for true belief (Section 4). Finally, I will suggest rough methods for carrying out a "beetle-box" measurement and make some concluding remarks (Section 5).

2. The problem

Dennett argues that, in general, it is "extremely difficult" to distinguish **B** from **BB**. To demonstrate the point, he constructs a hypothetical multiple-choice question:

God exists: __ Yes __ No __ I don't know

He then explains why this approach would fail:

*Submitted 12/19/08 in partial fulfillment of the course requirements for Philosophy 482 (professor Eric Lormand).

> Given the way religious concepts and practices have been designed, the very behaviors that would be clear evidence of belief in God are also behaviors that would be clear evidence of (only) belief in belief in God. If those who have doubts have been enjoined by their church to declare their belief in spite of their doubts, to *say the words* with as much conviction as they can muster, again and again, in hopes of kindling conviction, [. . .], then they will check the "Yes" box with alacrity, even though they don't really believe in God; they fervently believe in belief in God. (223)

Some unpacking needs to be done here, for there are really two claims. The first is that religious *behaviors* are, by design, untrustworthy guides to religious *beliefs*, at least as far as the **B**–**BB** distinction is concerned. That is, the kinds of things that religious people do—go to church, pray, celebrate all the big holidays, etc.—simply don't give enough information about the (first- or second-)orderedness of their beliefs.

I take this to be less a feature of *religious* behaviors and beliefs than it is of the distinction between beliefs and beliefs-in-belief in general. Consider the following scenario:

> Every student and teacher at Springfield elementary has the option to do ten hours of community service each year, or alternatively, to complete ten hours of certified meditation. Lisa, a student, believes that community service is fun, so she opts to work at a soup kitchen for two nights. Her teacher, Mrs. Krabappel, hates community service but believes in *the belief* that community service is fun—because she cares about the poor and would like more students to help them. So she decides to put in her ten hours at the soup kitchen after all, and raves about it afterward just as enthusiastically as Lisa.

Mrs. Krabappel is acting on a belief in belief—like a company that sells crucifixes—but here that belief is distinctly un-religious. And yet we have the same problem: hers and Lisa's behavior is basically indistinguishable, and if you surveyed them ("Is community service fun?") they would give the same answers.

(My example is not meant to show that Dennett *missed* something, but rather that his points are more general than they appear; he is writing a book *about* religion, so we would expect most of his discussion to be focused there.)

The second claim above is that those people who don't *actually* believe in God deliberately hide that fact, *because* they have been "enjoined by their church to declare their belief in spite of their doubts." (This, too, is not *just* a feature of religion—one can easily imagine principal Skinner enjoining all the teachers to do community service.) In other words, *pressure closes the gap between **B** and **BB***. We should keep that in mind for later.

But Dennett is quick to point out that it is not *just* that it is nigh impossible to distinguish **B** from **BB** by observing someone's behavior, and that people with **BB** have strong incentives—the fear of reprisal, their own social self-consciousness, etc.—to act *as though* they have **B**, but *also* that "you may find that when you look in your heart you simply do not know whether you *yourself* believe in God." The distinction between **B** and **BB** is not even articulable!

With this, it seems as though Dennett has dealt a fatal blow to the project of finding out who truly believes. But I would like to argue that there is still hope, if only we see that there are several degrees of freedom here, or "knobs" that we can turn, that in large part determine *just how hard* it will be to distinguish **B** from **BB**.

The two critical knobs are *articulability*, or the degree to which a person can examine and describe his beliefs, and *sharedness*, or the degree to which *my* "X" is the same as *your* "X." Another important parameter, and the last that we will consider, is *the kind of behavior that beliefs engender*. That will determine what we can observe and hence what we can (and should) test.

We will see that while in many configurations of these parameters the **B**/**BB** distinction remains totally buried, we can turn the knobs in ways that make it fall right out.

3. The beetle-box metric

After discussing the Druze—a peculiar religious community based in Beirut where residents insist on lying to outsiders about their beliefs—Dennett goes on to quote a passage at length from Wittgenstein's *Philosophical Investigations*:

> Suppose everyone had a box with something in it: we call it a "beetle." No one can look into anyone else's box, and everyone says he knows what a beetle is only by looking at *his* beetle.—Here it would be quite possible for everyone to have something different in his box. One might even imagine such a thing constantly changing. —But suppose the word "beetle" had a use in these people's language?—If so it would not be used as the name of a thing. The thing in the box has no place in the language-game at all; not even as a *something*: for the box might even be empty. —No, one can "divide through" by the thing in the box; it cancels out, whatever it is. [Section 293] (Dennett 235)

Dennett remarks that

> Much has been written on Wittgenstein's beetle box, but I don't know if anybody has ever proposed an application to religious belief. In any case, it seems fantastic at first that the Druze might be an actual example of the phenomenon.

Indeed, the idea of a religion with beliefs that cannot be observed and that (possibly) change constantly sounds a lot like Wittgenstein's hypothetical box. But the connection seems just as valid for any other religion, or any other belief at all. *Every* concept, from *God* to *I* to *chair*, is like a beetle in a box: we all use the same word "chair" and say we know what it means based only on our own personal, internal mental contents (the brain state we're in when we think of chairs), contents which constantly change.

There is a way in which some concepts seem *more* like beetles in a box than others, though. My concept of *Two*, for instance, is probably very much like everyone else's; we all have (roughly) the same beetle in our boxes. Thus our word "Two" is not pulling any tricks—it is not, as Wittgenstein puts it, that "the thing in the box has no place in the language-game," for the internal mental contents referred to by "Two" are (presumably) not arbitrary. Of all concepts, in fact, *Two* would probably be one of the *least* like a beetle in a box.

The word "I," on the other hand, and its corresponding concept of *me* or *my self*, is probably much closer to what Wittgenstein had in mind (and to the Druze's religion). Each person understands "I" based only on *his or her own self*, obviously, and every *self* is (just as obviously) different. But still, "I" has *a* place in the language-game, because everyone who says it is referring to the same *type* of object, even if the actual constitution of that object is unique. So it goes for any "relative reference": *that chair*, *the telephone closest to X*, etc. Example: "n is the biggest number I can think of" depends on who says it (and is thus just as relative as "I"), but since whoever says it is *doing the same sort of thing* (mentally) when he "processes" the phrase, it is not useless in the way that Wittgenstein's "beetle" is useless.

It should be no surprise, then, that religions—and in particular, their various conceptions of "God"—also admit to degrees of beetle-in-a-box-resemblance. As it happens, these are often distributed across time, with those *most* like a beetle in a box appearing *latest*. Dennett gives a run-through (though he doesn't realize he's taking steps up the beetle-box ladder): from rain gods and Greek gods to Yahweh of The Old Testament, the original New Testament Lord, "The genderless Person without a body who nevertheless answers prayers in real time (Stark's conscious supernatural *being*)," etc., all the way up to "a Higher Power (Stark's *essence*)."

At the far end are God-concepts that resemble *Two*, and at the close end are concepts that resemble *my self*. The distribution suggests what I will call the "beetle-box metric," which ranks God-concepts by their resemblance to Wittgenstein's "beetle," from lowest (least like the beetle) to highest (most like it).

Why is it useful to know whether a God-concept is more like *Two* than *me*? For one, if the historical correspondence above is more than just suggestive, it might be possible to date religions based on their beetle-box metric alone, where higher = more recent. That is far afield, though. More relevant to this paper is the idea that *the less of a "beetle in a box" a belief is, the easier it will be to distinguish **BB** from **B**.*

Consider the proposition that $4 + 7 = 11$. I can *believe* in that proposition (**B**) and *believe in believe of it* (**BB**), or I can have **BB** *without* **B**. Although this latter option is absurdly unlikely for anyone *normal*, we can easily imagine a group of "weirdos" who (a) do arithmetic in base 8, and so believe that $4 + 7 = 13$ but (b) believe in belief that $4 + 7 = 11$, because they realizes that society would go to shambles if people stopped using base 10.

The point is, picking out a weirdo would be trivially easy, as long he wasn't self-conscious about his weirdness: one could simply ask "What is four plus seven?" and he should answer "thirteen" while everyone else says "eleven." The two critical points are (a) each weirdo's concept of arithmetic is readily articulable, viz., he never has to ask himself "which base am I working in here?", and (b) all weirdos act alike, because they all understand the same by the question (since they all do base 8 arithmetic).

The analogy back to religion should be apparent. Beetle-box metric in hand, we can begin classifying religious beliefs and, more usefully, devising tests that distinguish **B** from **BB** in each of the classes.

4. Four broad classes

One can imagine turning our *articulability* and *sharedness* knobs arbitrarily, haphazardly exploring the beetle-box space (and the space of possible religions). It will be more perspicacious, I think, to divide it instead into four classes, and leave all that knob-turning for some other time. These four classes are:

1. Everyone has roughly the same concept X in their box, and X is (2. is not) articulable.
3. Everyone has a different concept X in their box, and X is (4. is not) articulable.

Shared, articulable

The concept *Two* given above is a member of this class. Most concepts are, in fact. I think there is a good (if obvious) reason for this: a language where most words had arbitrary referents (i.e., where most words were like Wittgenstein's "beetle") would be terrible for communication—nobody would know what anyone else was grunting about. Further, concepts that are not articulable are, by design, difficult to put into words, and they would likely appear last in a language built up by trial and error.

One's reasons for believing in belief in a shared, articulable concept X should be just as concrete and apparent as one's reasons for *believing* X: one knows what belief in X entails (because it is articulable), and since X is shared, one knows what belief in belief in X entails (by simply imagining one's own belief multiplied through the community). Thus in the absence of social pressure to *profess* belief, a simple questionnaire should reveal whether someone *actually* believes in X or whether they only believe that they (or others) *should* believe in X. If there *is* pressure, it should be sufficient to observe someone's private behavior, because they would have no reason to act as though they believed in X if they merely believed in belief in X, and there would be no confusion about which is which.

Shared, non-articulable

Following Hilary Putnam's "division of linguistic labor," Dennett argues for a kind of "division of *doxastic* labor," whereby the "work" of actually understanding a commonly held belief X is reserved for a few experts.

One example of his is Einstein's familiar equation $e = mc^2$. He argues that in cases like this that we are "not really *believing the proposition*. For that, you'd have to *understand* the proposition. What we are

doing is believing that *whatever proposition is expressed by the formula 'e = mc² ' is true.*" In other words, we do not really *believe* that $e = mc^2$; we believe that we *should* believe that $e = mc^2$—because an expert *actually* believes it.

Since belief here is tantamount to understanding, one easy test for it is to have an expert[1] interrogate the subject; if (and only if) he discovers that the subject understands the concept as well as himself, then the subject can be said to believe it. Otherwise, the best he can get is belief in belief.[2]

Different, articulable

Everyone has an intricate concept of themselves that they could describe at length, yet everyone's *I* is different—thus *I* is a member of this class. In the same vein, most everything *reflexive* belongs here, as do all manner of *preferences* (e.g. favorite color).

For an example of **BB** without **B**, imagine that a person (a) believes he is hideous, but (b) believes that everyone (including himself) *should* believe they're beautiful. Such a person would probably answer "Yes" to the naïve question, "Do you think you're beautiful?", because they believe in positive beliefs about themselves (a positive self-image).

The implicit suggestion here is that since concepts in this class are articulable, someone who believes in belief in an *X* but does *not* believe in *X* will *know the difference*—they will in principle be able to answer for themselves the question "Do I really believe *X*?" So anyone who *claims* to believe in *X* while actually *not* believing it (perhaps they merely believe in belief in *X*) is knowingly deceiving the questioner.

That feature leads us into two possible tests: the first, call it "the Dr. Phil," seeks to coax the truth out of a subject in the same way that a skilled psychologist might: by asking sympathetic, leading questions. The second, call it "the Jack Bauer," would have someone forcefully obtaining the truth based on the principle that a subject will "crack" if only pushed hard enough. Both rely on the fact that the subject knows she is withholding the whole truth.

Different, non-articulable

This is the land of wonky concepts: *the last digit of* π, for instance. It is not articulable, in that I do not (and cannot) have mental access to it; the best I can offer is that it is not zero (but I even have my doubts about that). Nor is it shared—for one, the digit itself could vary across people (John could think *4* and Bill could think *9*), but so too could the *idea*: for some it might mean "a number between zero and nine," and for others "the unknowable" or "an absurdity" or "the eye of God."[3]

There is nothing insidious about one of these concepts on its own—if anything, it passes the time and exercises one's imagination. But Dennett's warning, that

> Once people start committing themselves (in public, or just in their "hearts") to particular ideas, a strange dynamic process is brought into being, in which the original commitment gets buried in pearly layers of defensive reaction and meta-reaction,

[1] Such an expert is not guaranteed to exist, of course. Consider the first prime number *N* greater than the greatest prime yet discovered. *N* is not articulable to *anybody* (yet), but everyone who utters the phrase "the first prime number *N* greater than the greatest prime yet discovered" is referring to precisely the same thing—everyone has the same beetle in their box, even though no one can see it. Concepts of this type are not all that pertinent to the current discussion, though, since no one can believe them.

[2] Can someone believe in a shared, non-articulable concept and not believe in belief of it? Sure: imagine a common misconception, say, that shaving causes hair to grow back thicker and coarser. No one except experts understands it well (non-articulable), and everyone means the same thing when they refer to it (shared). Then imagine that you yourself believe it, but, because of your lack of training in biology and your abundance of training in cognitive bias, you actively avoid spreading the idea.

[3] These are articulations, to be sure, but they are not articulations of *the last digit of* π, but rather of "the last digit of π." They are, in other words, articulations of "what we would be looking for if we were searching for *the last digit of* π," and not of the actual object of our search.

is most apt here, when one's "particular ideas" are formless and private—because those ideas act like "wildcards." That is, ideas that have not been articulated (much) are not yet committed to (m)any facts, and so are compatible with (m)any *arbitrary* fact(s); moreover, ideas that are private (not shared) cannot, in principle, undergo the kind of "compatibility checking" with an expert that was possible in the *Shared, non-articulable* class.

The trouble, then, is that it is *easy* to maintain one's commitment these "wildcard" ideas, because there is no inconsistency—logically, cognitively, or publicly—in changing their content if the commitment so demands it. What then happens, as Dennett puts it, is that whatever little actual articulable content comprises the idea gets buried under these changes, or attempts to attack and defend it (his "pearly layers"). That is far less likely when an idea is shared—because an expert's articulation (the "orthodoxy") is available—or articulable—because there are more committed-to facts to fix an idea in place.

So someone who claims true belief, but who actually only has a particularly powerful *belief in* (or commitment to) belief, could plausibly *not know it*, because the truth is buried under layers of cognitive infighting. They can actually convince *themselves* that they have **B** when they really only have **BB**.

Thus neither "the Dr. Phil" nor even "the Jack Bauer" is likely to distinguish **B** from **BB** in this class, because the subject is not *knowingly* withholding their true beliefs. So we need another test.

In *Indiana Jones and the Last Crusade*, Harrison Ford's character (Indy), before he can enter the "grail room" that contains the cup of Christ, must overcome three challenges. The first of these is a set of blades that one can avoid, legend has it, by being "humble before God"—i.e., by kneeling. Next is a floor divided into letters, which Indy infers—based on the clues "only in the footsteps of God will he proceed" and "word of God"—must be stepped on in the precise order I-E-H-O-V-A. The last is a simply a large chasm, which leads Indy to believe he must take a "leap of faith" to pass.

This final challenge is special because, at least in Indy's mind, there is no *trick* to it—like there was when "humble before God" *actually* meant "bend down" and "only in the footsteps of God will he proceed" *actually* meant "only by walking on tiles in a certain order will he not fall through." He thinks that he must *truly* believe to survive.[4]

Without "brain reading" technology, such a test seems like the best method for penetrating those "pearly layers of defense reaction and meta-reaction" that shroud unbelief in belief in belief.[5] One would expect someone who *actually* believes to take the leap, and those who don't to back away, kick sand into the chasm (in the hopes of revealing a hidden bridge), or simply break down. Of course it is plausible that someone's self-conditioning could be so strong that they take the leap in the absence of true belief, which would break our test, but I think we must tolerate such false positives until we find something better.

5. Classifying religions

We can now use our beetle-box metric to divide religious beliefs and God-concepts into our four broad classes, which will help us determine the appropriate **B**/**BB** test. I will give a few (simplified) examples, and then discuss some of the general tools one can use to classify religions (according to the beetle-box metric) in general.

1. *Roman Catholics.* Catholics (more so than Protestants) are said to have elaborate and precise rituals, as well as a long history of religious scholarship interpreting the Bible. While many adherents probably waver from the orthodoxy, let us assume that they have all read and internalized the clergy's

[4]Only after he takes his leap does he realize that there actually *was* a trick, viz., that there was bridge across the chasm all along, but that it was made virtually invisible; he didn't need *actually* need true belief, after all. The example works, though, because he *believed* he did.

[5]It is not *just* that you are putting someone's life on the line. Jack Bauer could do that, too. It is that the subject believes that *God* is testing them.

articulations of the religion's traditions, practices, and beliefs. Thus we would place much of Roman Catholicism into the *shared, articulable* class.

2. *New Life Church.* Campus ministries often attract college students with loose, metaphorical interpretations of the Bible; a focus on "real-life issues," like drinking, politics, and sex; a Higher Power type of God; and an overwhelming emphasis on faith (belief in belief). One is tempted to place these religious concepts into the *different, non-articulable* class.

3. *Buddhism.* Presumably[6], "Buddhists" who meditate regularly and who nurture a personal, inarticulable spirituality, *also* ascribe to the Way, the eight-fold path, *samsara*, and the like—shared concepts in all of Buddhism. This would tempt me to place their beliefs, and in particular their "God"-concept (interpreted loosely) into the *shared, non-articulable* class.

4. *Personal God.* I can imagine that many people harbor a "personal God," to whom they pray and of whom they have a relatively stable image: if you asked them to describe It, assuming they weren't embarrassed, they might give you some features of Its personality, including a list of things that seem to "set It off" (corresponding, I would venture, to deeds with negative worldly consequences), a vague physical description, and an account of Its warmth and guidance. Such personal Gods would seem to occupy the *different, articulable* class.

In general, it should be possible to get a sense of a religion's beetle-box class. For one, it seems (based on Dennett's account) that *shared, non-articulable* God concepts are a relatively recent historical invention. As one proceeds backward in time, the God concepts get correspondingly more concrete, to the point of folk religions whose tenets seemed to be mostly *shared* and *articulable*. Thus one simple method for classifying an unknown religion would be to acquire a rough estimate of its origin in time.

An abundance of holy texts should generally be a good sign of *shared* ideas, or doxa; an emphasis on free-wheeling debate or discussion, as opposed to structured (e.g. catechismal) discourse, would be another. Articulability is readily assessed via interviews or conversations with adherents.

Indeed, it would seem a worthy project to design a survey that might help one classify religions by our beetle-box metric or one more fine-grained. With that done, the task of distinguishing true believers would be made more manageable.

Conclusion

I hope I have given a reasonably good method for classifying religious concepts both for its own sake and, if one applies the appropriate tests, to help distinguish **B** from **BB**. Further work might refine these classes, develop empirical means for deciding which religions belong in which classes, and suggest more (and more precise) tests for distinguishing true belief.

[6] I suspect that this is one of (many of) those areas where I have just enough knowledge to be dangerous, and not much more.