Dissolving the Question

by **Eliezer Yudkowsky** 11y 109 comments

50

"If a tree falls in the forest, but no one hears it, does it make a sound?"

I didn't *answer* that question. I didn't pick a position, "Yes!" or "No!", and defend it. Instead I went off and deconstructed the human algorithm for processing words, even going so far as to sketch an illustration of a neural network. At the end, I hope, there was no question left—not even the feeling of a question.

Many philosophers—particularly amateur philosophers, and ancient philosophers—share a dangerous instinct: If you give them a question, they try to answer it.

Like, say, "Do we have free will?"

The dangerous instinct of philosophy is to marshal the arguments in favor, and marshal the arguments against, and weigh them up, and publish them in a prestigious journal of philosophy, and so finally conclude: "Yes, we must have free will," or "No, we cannot possibly have free will."

Some philosophers are wise enough to recall the warning that most philosophical disputes are really disputes over the meaning of a word, or confusions generated by using different meanings for the same word in different places. So they try to define very precisely what they mean by "free will", and then ask again, "Do we have free will? Yes or no?"

A philosopher wiser yet, may suspect that the confusion about "free will" shows the notion itself is flawed. So they pursue the Traditional Rationalist course: They argue that "free will" is inherently self-contradictory, or meaningless because it has no testable consequences. And then they publish these devastating observations in a prestigious philosophy journal.

But *proving that* you are confused may not make you feel any *less* confused. Proving that a question is meaningless may not help you any more than answering it.

The philosopher's instinct is to find the most defensible position, publish it, and move on. But the "naive" view, the instinctive view, is a fact about human psychology. You can prove that free will is impossible until the Sun goes cold, but this leaves an unexplained fact of cognitive science: If free will doesn't exist, what goes on inside the head of a human being who thinks it does? This is not a rhetorical question!

It is a fact about human psychology that people think they have free will. Finding a more defensible *philosophical position* doesn't change, or explain, that *psychological fact.* Philosophy may lead you to *reject* the concept, but rejecting a concept is not the same as understanding the cognitive algorithms behind it.

You could look at the Standard Dispute over "If a tree falls in the forest, and no one hears it, does it make a sound?", and you could do the Traditional Rationalist thing: Observe that the two don't disagree on any point of anticipated experience, and triumphantly declare the argument pointless. That happens to be correct in this particular case; but, as *a question of cognitive science,* why did the arguers make that mistake in the first place?

The key idea of the heuristics and biases program is that the *mistakes* we make, often reveal far more about our underlying cognitive algorithms than our correct answers. So (I asked myself, once upon a time) what kind of mind design corresponds to the mistake of arguing about trees falling in deserted forests?

The cognitive algorithms we use, *are* the way the world feels. And these cognitive algorithms may not have a one-to-one correspondence with reality—not even macroscopic reality, to say nothing of the true quarks. There can be things in the mind that cut skew to the world.

For example, there can be a dangling unit in the center of a neural network, which does not correspond to any real thing, or any real property of any real thing, existent anywhere in the real world. This dangling unit is often useful as a shortcut in computation, which is why we have them. (Metaphorically speaking. Human neurobiology is surely far more complex.)

This dangling unit *feels like* an unresolved question, even after every answerable query is answered. No matter how much anyone proves to you that no difference of anticipated experience depends on the question, you're left wondering: "But does the falling tree *really* make a sound, or not?"

But once you understand *in detail* how your brain generates the *feeling* of the question—once you realize that your feeling of an unanswered question, corresponds to an illusory central unit wanting to know whether it should fire, even after all the edge units are clamped at known values—or better yet, you understand the technical workings of Naive Bayes—*then* you're done. Then there's no lingering feeling of confusion, no vague sense of dissatisfaction.

If there is *any* lingering feeling of a remaining unanswered question, or of having been fast-talked into something, then this is a sign that you have not dissolved the question. A vague dissatisfaction should be as much warning as a shout. *Really* dissolving the question doesn't leave anything behind.

A triumphant thundering refutation of free will, an absolutely unarguable proof that free will cannot exist, feels very *satisfying*—a grand cheer for the home team.   And so you may not notice that—as a point of cognitive science—you do not have a full and satisfactory descriptive explanation of how each intuitive sensation arises, point by point.

You may not even want to admit your ignorance, of this point of cognitive science, because that would feel like a score against Your Team. In the midst of smashing all foolish beliefs of free will, it would seem like a concession to the opposing side to concede that you've left anything unexplained.

And so, perhaps, you'll come up with a just-so evolutionary-psychological argument that hunter-gatherers who believed in free will, were more likely to take a positive outlook on life, and so outreproduce other hunter-gatherers—to give one example of a completely bogus explanation. If you say this, you are *arguing that* the brain generates an illusion of free will—but you are not *explaining how.* You are trying to dismiss the opposition by deconstructing its motives—but in the story you tell, the illusion of free will is a brute fact. You have not taken the illusion apart to see the wheels and gears.

Imagine that in the Standard Dispute about a tree falling in a deserted forest, you first prove that no difference of anticipation exists, and then go on to hypothesize, "But perhaps people who said that arguments were meaningless were viewed as having conceded, and so lost social status, so now we have an instinct to argue about the meanings of words." That's *arguing that* or *explaining why* a confusion exists. Now look at the neural network structure in Feel the Meaning. That's *explaining how*, disassembling the confusion into smaller pieces which are not themselves confusing. See the difference?

Coming up with good hypotheses about cognitive algorithms (or even hypotheses that hold together for half a second) is a good deal harder than just refuting a philosophical confusion. Indeed, it is an entirely different art. Bear this in mind, and you should feel less embarrassed to say, "I know that what you say can't possibly be true, and I can prove it. But I cannot write out a flowchart which shows how your brain makes the mistake, so I'm not done yet, and will continue investigating."

I say all this, because it sometimes seems to me that at least 20% of the real-world effectiveness of a skilled rationalist comes from not stopping too early. If you keep asking questions, you'll get to your destination eventually. If you decide too early that you've found an answer, you won't.

The challenge, above all, is to notice when you are confused—even if it just feels like a little tiny bit of confusion—and even if there's someone standing across from you, *insisting* that humans have free will, and *smirking* at you, and the fact that you don't know *exactly* how the cognitive algorithms work, has *nothing to do* with the searing folly of their position...

But when you can lay out the cognitive algorithm in sufficient detail that you can walk through the thought process, step by step, and describe how each intuitive perception arises—decompose the confusion into smaller pieces not themselves confusing—*then* you're done.

So be warned that you may *believe* you're done, when all you have is a mere triumphant refutation of a mistake.

But when you're *really* done, you'll *know* you're done. Dissolving the question is an unmistakable feeling—once you experience it, and, having experienced it, resolve not to be fooled again. Those who dream do not know they dream, but when you wake you know you are awake.

Which is to say: When you're done, you'll know you're done, but unfortunately the reverse implication does not hold.

So here's your homework problem: What kind of cognitive algorithm, as felt from the inside, would generate the observed debate about "free will"?

Your assignment is not to argue about whether people have free will, or not.

Your assignment is not to argue that free will is compatible with determinism, or not.

Your assignment is not to argue that the question is ill-posed, or that the concept is self-contradictory, or that it has no testable consequences.

You are not asked to invent an evolutionary explanation of how people who believed in free will would have reproduced; nor an account of how the concept of free will seems suspiciously congruent with bias X. Such are mere attempts to *explain why* people believe in "free will", not *explain how.*

Your homework assignment is to write a stack trace of the internal algorithms of the human mind as they produce the intuitions that power the whole damn philosophical argument.

This is one of the first real challenges I tried as an aspiring rationalist, once upon a time. One of the easier conundrums, relatively speaking. May it serve you likewise.

50

Next:

Wrong Questions

127 comments 41 points

Eliezer Yudkowsky's commenting guidelines: Reign of Terror - I delete anything I judge to be annoying or counterproductive

Frontpage commenting guidelines:

Aim to explain, not persuade. Write your true reasons for believing... (Read More)

109 comments, sorted by oldest Highlighting new comments since Today at 5:09 PM

Please log in to comment.

[-] **Tom_McCabe2** 11y 5

I have no idea why or how someone first thought up this question. People ask each other silly questions all the time, and I don't think very much effort has gone into discovering how people invent them.

However, note that most of the silly questions people ask have either quietly gone away, or have been printed in children's books to quiet their curiosity. This type of question- along with many additional errors in rationality- seems to attract people. It gets asked over and over again, from generation unto generation, without any obvious, conclusive results.

The answer to most questions is eith... (Read more)

[-] **Maksym_Taran** 11y 4

I think a brain architecture/algorithm that would debate about free will would have been adapted for large amounts of social interaction in its daily life. This interaction would use markedly different skills (eg language) from those of more mundane activities. More importantly it would require a