

# Self-Belief

By Paul Almond, 10 June 2011, <http://www.paul-almond.com>, [info@paul-almond.com](mailto:info@paul-almond.com)

## Introduction

This essay is about a thought experiment that I found interesting enough to present for discussion. In the thought experiment, a light will turn on if you believe it will turn on and will not turn on if you do not believe it will turn on. If the light turns on, you will win a prize. The interesting feature of this is that both belief and non-belief turn out to be correct, yet clearly you should want to believe that the light will turn on and win the money. How easy is it to do this, and how exactly do you do it? What "ability" is being tested by this? I think that this issue merits more discussion than this essay will give it.

## The Scenario

You are sitting at a table. There is a light on the table. The light is currently turned off. A "mind reading machine" is set up to analyze your brain activity continually. (Imagine if you like that it does this by picking up brainwaves through electrodes: the exact mechanism doesn't matter).

When I say, "Go," the experiment will start. The mind reading machine will analyze your brain activity for one minute. (Change this time if you like.) When the minute is up, the mind reading machine will either turn the light on, or it will leave it turned off according to a rule of which you are fully aware. (Nothing about the setup here is hidden from you.) If the light turns on, you win £1,000,000. (This is a thought experiment: I am not really offering anyone any money!)

This is the rule:

- If the mind reading machine determined that you tended more towards believing that the light *would* turn on, for the minute it analyzed you, when the minute is up, *it will turn the light on*.
- If the mind reading machine determined that the above situation did not occur - that you did *not* tend more towards believing that the light would turn on - it will *not* turn on the light.

So, the light turns on after one minute if you believed it was going to do that. If you did not believe it was going to turn on, it does not turn on. (Of course, belief may not be an all or nothing thing - which is why I said "tended more towards". However you think belief works, you should be able to think of this in some terms such that some kind of mechanism which measures your degree of belief one way or the other is possible in principle.)

### **Both belief and non-belief turn out to be correct.**

Interestingly, everyone who goes through this experiment is shown to be right. If Fred spends one minute thinking that it is absurd that the light will turn on, the machine will record this, and it will not turn on: Fred's belief will turn out to be correct. If Jane spends one minute thinking that the light will definitely turn on, it will. Jane's belief is similarly correct. But this does not mean Fred and Jane are both being rational. They may each have silly reasons for their beliefs: even if you believe that the light will turn on because toasters are eating liberalism, you will appear to get vindication.

There is also a kind of recursive nature to the whole thing. If Fred fails to win the money, he may see Jane walking out with the money and demand another try, but what is he supposed to do to make himself believe? He knows that if he believes he will win the money, but is he not being asked here to believe that he will believe? What if he does not believe that he will believe? Well, he could win the money if he can believe that he will believe that he will believe, and so on, but where is there any “grounding” for any of this? How is Fred supposed to get this grounding?

### **Would you win the money?**

The obvious question is: if we did this with *you*, do you think you would win the money (the light turns on) or not win it (the light does not turn on) and why? I will say that most people who have expressed an opinion on this so far seem to think they would have no trouble winning the money: they seem to think they would easily believe that the light will turn on for them.

It seems that a lot of people would believe that they could win the money because “they believe in themselves”. Now, if people think that I can hardly argue against it. By definition, if someone thinks he is going to win the money, and really thinks it, he will win the money. This raises the issue here, though, of what exactly such a person is doing to win the money. Why should he get that belief in the first place? If you do not have that belief, it seems desirable to get it – if you want the money – but how do you go about getting the belief? Are you supposed to practice? What would you say to someone, desperate to win the money, and who has had tried to win the money many times, each time failing to believe appropriately? Would you tell him that he needs to believe? How is he supposed to do that? All the previous evidence should suggest that he will not believe strongly enough – meaning that it should cause him not to believe strongly enough, and yet if he could just make himself believe, his fortunes would change immediately.

### **Is the scenario circular?**

There may *appear* to be a circular nature to the whole situation, yet there cannot be really. The situation may also seem to have some similarity with Newcomb's paradox (Nozick, 1969; Kiekeben, 1996), but there should not be the same controversy about whether the situation is coherent or possible that there is with Newcomb's paradox. If we

had the technology we could probably build a machine like this. No laws of science seem violated by it: we could run this experiment for real if we had the technology. In fact, I expect that someone will one day. We could even run a crude version of it now, for example by using a lie detector and asking the subject, at the end of the minute, if he tended to believe that the light was going to turn on during the minute. You could also try a “dry run” of this by sitting there for a minute, imagining that you are in the experiment, and trying to imagine what you would be thinking. Some readers have probably already done that by the time they have read this far. Unlike with Newcomb’s paradox there can be no reasonable controversy about what you *should* do: it is quite obvious that you should believe that the light will turn on. The issue is about what you *would* do – and that question is more personal in nature.

### **Is the scenario like Pascal’s wager?**

The scenario here might also be considered similar to Pascal’s wager, which suggests that you should believe in God in case he exists, so that you “win” eternal life if he does. Any such resemblance is really superficial, though. Pascal’s wager is asking you to believe something irrespective of the likelihood that it is true. With this scenario, if you can make yourself believe, your belief will automatically become correct: you cannot be “wrong”.

### **One Idea**

One way of approaching this – and this is purely speculative, as I am mainly offering this for discussion rather than trying to solve anything – may be to say that belief can be thought of as going one way or another like a ball rolling off the top of a ridge, so at the start it is easy for it to get pulled one way or the other, and as it gets less neutral it becomes less likely that it will be pulled back: the previous slight tendency to believe that one result or the other will occur becomes evidence for still stronger belief, and so on.

Many readers will recognize this as a “chaos theory” type view, in which the belief ends up in the vicinity of one attractor or another. Someone who started with a small tendency towards thinking that the light would come on might justifiably think that this puts him at an advantage, as his existing belief is evidence of belief, so that strengthens the belief that the light will come on still further, and so on. Likewise, someone who starts with a small tendency towards disbelief might find it hard to escape from this.

This suggests that, maybe, for a rational person, it may be that the initial “random noise” that is there before a proper belief gets established might determine which way things are likely to go, and as they go further one way they are less likely to come back, but someone might start going through all this before even sitting down at the table. Someone who has to play this game every day for example, would have more time to develop a belief, and would know what had happened in previous runs. We might, however, think that there may be a capability to “choose” the belief – and this would only need to be very limited. An individual would only need to be able to “nudge” his belief very slightly “off the ridge” towards believing that the light will come on – either by “choosing” to do

so, or by some tendency towards “self-belief” or “optimism” – and then this belief would itself be evidence for stronger belief.

We might also consider what the uses, if any, would be of the kind of apparatus used in this thought experiment, if it could be made in reality. Would practicing turning on the light have any use in other respects? Would it be practice of some “self-belief” or “optimistic belief” ability that could have more general use? Could someone be trained to turn on the light as a way of training them to be better at adopting unreasonable beliefs?

## References

Kiekeben, F., 1996. *Newcomb's Paradox*. [Online] Franz Kiekeben's Page. Available at: <http://www.kiekeben.com/newcomb.html> [Accessed 10 June 2011].

Nozick, R., 1969. Newcomb's Problem and Two principles of Choice. In: Rescher, N. et al, eds., 1969. *Essays in Honor of Carl G. Hempel*. Dordrecht: D. Reidel. pp.114-115. Reprinted in Campbell, R. & Sowden, L., eds., 1985. *Paradoxes of Rationality and Cooperation*. Vancouver: University of British Columbia Press.