

# Re: Halting Problem for Humans

---

*Source:* <http://sci.tech-archive.net/Archive/sci.logic/2006-10/msg00652.html>

---

- *From:* "LauLuna" <[laureanoluna@xxxxxxxx](mailto:laureanoluna@xxxxxxxx)>
  - *Date:* 23 Oct 2006 10:38:39 -0700
- 

They are completely well-defined questions. Either Peter's response will

be "no", or it won't. Either Daryl's response will be "no" or it won't. There is no third possibility.

Yes, that's right. I should have reflected on it more carefully. Please, consider the following.

If you want to prove anything about limits of human knowledge, you have to suppose your proof goes through for the the best possible human thinkers and knowers.

So suppose Peter and Daryl are ideal thinkers and knowers; then each of them knows they are. We do not need to suppose they are honest and observe the rules of the game when answering; it suffices if each of them knows that the answer of the other will be in some way determined by his attempt to know. The situation that they are honest and they both know they both are is a particular case of my approach.

Each of them will try to figure out what the other will think, but since each of them knows the other will do the corresponding, each of them will be compelled to consider his own attempt to answer while performing that attempt.

So, ideal thinkers would find themselves trapped in circularity. And knowing the other is also trapped won't help them. None of them will ever work out what the other is going to say because it would imply taking into account his own thinking in order to decide what his own thinking should be.

Exactly this I said was phenomenologically impossible. So, what I thought worked against your proposal actually works for it.

Now suppose one (or both) of our ideal thinkers does not observe the rules of the game and answers even if he doesn't know the correct answer. Well, that won't change the fact that no one knows what the

## Re: Halting Problem for Humans

other will answer. Fortuitous success is here of no interest, I think.

By splitting into a two isolated person game the single person game you have produced a beautiful piece where the 'escape' clause seems out of order. It seems there is a possible situation in which ideal knowers won't be able to know. I do think your example shows something interesting.

I believe circularity lies in the core of the halting problem, for humans and for machines. But I think it is quite different for humans; for humans it is a question of 'intentional circularity': for humans the relevant fact is that no intentional act can be its own intentional object. E. g. when I'm thinking that it will rain I cannot be thinking at the same time that I'm thinking it will rain. It is easy to see that if an intentional act were its own intentional object, that object would be of infinite complexity.

I think you have shown that an ideal thinker is not just incapable of always having his own thought as an object but also incapable of always having as an object the thought of other ideal thinker: no one can set his thought at a higher level than the other's, as no one can set his thought at a higher level than itself.

Now, there is neither logical nor scientific proof that human behavior is only determined by physical causality and that mental states are causally inert. This can only be a philosophical (or ideological) assumption.

Suppose the ideal attempt of our ideal thinkers could be represented as a physical process completely determined by physical causes. Then, if each of them is provided with a computer that has previously collected sufficient information about the other's brain (and a relevant part of its environment) and can calculate physical causality, each of them should be able to know what the other would say.

Nevertheless, this seems impossible: what would prevent them from answering what they know?

Regards

Daryl McCullough wrote:

LauLuna says...

There is a crucial difference between the case you propose and the halting problem. The halting problem is a well defined question while

## Re: Halting Problem for Humans

your questions to Peter and Daryl are not.

Not really. Imagine that Peter and Daryl are both robots, and their behavior is completely specified by a computer programs, and that both Peter and Daryl have access to each other's programs. In that case, it is exactly analogous to the halting problem.

In the case of real humans, it only becomes fuzzier because we lack perfect knowledge about the mechanisms by which each of us makes decisions.

Each of them is circularly defined. Trying to complete the text of the question to Peter (which I suppose you think contextually completed)

The question was already complete. Peter was asked whether Daryl's next utterance will be "yes". Daryl was asked whether Peter's next utterance will be "no". That's a complete question.

we would obtain:

'Will Daryl answer 'yes' to the question whether you will answer 'no' to the question whether Daryl will answer 'yes' to the question whether...?'

Yes, you can expand any question into an equivalent longer question. The point is, there is no ambiguity about what each is asked to do.

'Will Daryl answer 'yes' to the question whether you will answer 'no' to this question?'

Now the self-reference only makes the circularity still more evident.

What difference does it make whether it is self-referential or not? There is no ambiguity in what each is being asked to do.

If they were completely described by computer programs, then it is clear what is being asked:

Daryl is being asked: Will Peter's response to string S be "no"?  
Peter is being asked: Will Daryl's response to string S be "yes"?

where string S is just the description of the game. There is no uncertainty or ambiguity here, other than the fact that (in the

## Re: Halting Problem for Humans

case of humans) we each only have partial knowledge of the behavior of others.

The halting problem contains no circularity in its definition because it poses a purely syntactical or mechanical question with no semantic or intentional dimension.

There is no semantic dimension here, either. Human beings are physical systems that work through natural laws. If we had perfect understanding of those laws (and of the initial conditions), then we would know what Peter's response will be (or else, his response could be nondeterministic, in which case we could make probabilistic predictions).

In contrast, and using phenomenological terms, we can say that no intentional act of thinking can be its own intentional object; call 'PNS' that proposition; so according to PNS, it is not the case that any intentional act is a possible object for any thinking subject.

I'm not sure what you are talking about. The question of what Peter's response to the game will be is a purely physical question. There is a causal chain from the act of Peter being told the rules to Peter producing certain sounds from his mouth. We only have partial knowledge of that causal chain, but I don't see how in principle there is any difference between asking a question about the response of a human being and asking a question about the behavior of a computer program.

This is why your questions to Peter and Daryl cannot be well defined questions for none of them:

They are completely well-defined questions. Either Peter's response will be "no", or it won't. Either Daryl's response will be "no" or it won't. There is no third possibility. (Well, actually, the third possibility is that they will not answer, or will answer using an illegal reply such as "I don't know"—but in that case, it is false that the answer is "yes" and it is also false that the answer is "no").

trying to answer them correctly would force them to consider their own attempt to answer and this would be phenomenologically impossible.

No, that's not correct. For Daryl to answer the question, he only needs to know how \*Peter's\* mind works. It's not self-referential in that respect. Daryl only needs to know what Peter \*believes\*.

Re: Halting Problem for Humans

Those beliefs may not be correct.

Maybe we've met here an essential difference between thinking behaviors and machines: only the latter are always possible objects for the former.

I think that's incorrect.

—

Daryl McCullough  
Ithaca, NY