Theory of mind and reasoning complexity

1 Introduction

In her useful review on theory of mind (TOM), Miller (2006) indirectly defines TOM as "An appreciation of others' thoughts, feelings, knowledge, and wishes". Malle in his review (2002) gives the more formal definition "the ability to represent, conceptualize, and reason about mental states".

Both authors focus on the connections between TOM and language, mainly concerning themselves with developmental psychology, clinical psychology and evolutionary psychology. But there are many other connections one might want to explore.

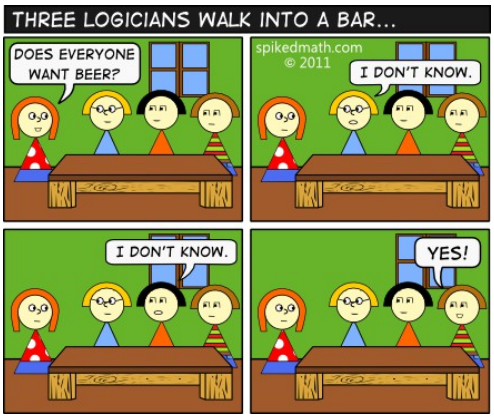
First, from a psychometric point of view, one may note that the TOM tests (review in Miller (2006: 150)) are similar to other developmental tests such as the Piagetian tests (Jensen 1980: 669-676). All such tests have a simple pass/fail scoring and as children age, they learn to pass tests that they would have failed earlier.

Second, from a comparative animal psychological perspective, one may note that some species do well on TOM tests while others don't. Dogs seem to do well while nonhuman primates don't (Hare et al. 2005).

Third, from a pragmatic perspective, one may note the interplay of TOM and conversational implicature (Horn 1989; Davis 2013).

However, in this paper I want to focus on the complexity of the TOM reasoning, by analyzing a particular example of it with formal logic.

2 Three logicians walk into a bar...



Consider the webcomic to the right.[1] I will bet that many if not most people can work it out in non-formal terms if they are given a bit of time. But as Miller (2006: 243) notes, TOM reasoning "is complex when made explicit, yet we do it all the time, with little or no conscious reflection". This means that TOM is normally a system I (intuitive, automatic, fast) process as opposed to a system II process (conscious, manual, slow), see Kahneman (2011).

2.1 The analysis

The analysis requires an understanding of formal logic. I have tried to make it as simple as possible for readers unfamiliar with the topic, as well as included translations of formalizations.

Let "E" refer to the proposition expressed by the sentence:

"Everyone wants beer"

"everyone" refers to the three people on the right in the webcomic.[2] Let's call them "a", "b", "c" from left to right. Let "Wx" =df "x wants beer".

We could try to show this but let's just take it intuitively that the following holds:

Everyone wants beer ↔ a wants beer and b wants beer and c wants beer

Formalizing the above we get:

$E \leftrightarrow (Wa \wedge Wb \wedge Wc)$

Now, assume that to begin with, a, b, and c does not know whether the two others want beer or not. This is technically 'left open' in the comic, but it is not irrelevant.

Now, assume every person knows if he wants beer or not, or rather, he either knows that he

1 The following analysis originally appeared on my blog, see Kirkegaard (2011).

2 One can complicate things further by including the questioner in the "everyone" category. If he is like the others, then he wants beer to begin with, otherwise he would already know the answer to his own question.

wants beer, or he knows that he does not want beer. Without this assumption about knowledge of one's own mental desire, it doesn't work either. Introducing "Kx(P)" to mean "x knows that P" and formalizing:

$(\forall x)Kx(Wx) \vee Kx(\neg Wx)$

[For any x, either x knows that x wants beer, or x knows that x does not want beer.]

Now, interpreting "logicians" to be a group of people being perfect at making inferences in at least this case. Such people are sometimes called "ideally rational" or similar. They never make a wrong inference and never miss an inference. For a review of different kinds of rational agents see (Wikipedia; Smullyan 1986).

Let's think about a's position as he is first to answer the question. If he knows that he wants beer (Ka(Wa)), then he does not know the truth of E. But if he knows that he does not want beer, he can know that E is false. Because: he is part of everyone, so if he does not want beer, it is not the case that everyone wants beer. If a is being truthful, he has to answer either "I don't know" or "No".

Now b is in almost the same position as a. Obviously, if a has already answered "No", there is no reason to respond or at least he should respond the same. If a's answer is "I don't know", b can infer that a wants beer (since, he should have said "No" if he didn't want beer). However, b still lacks information about whether or not c wants beer [Wc or ¬Wc]. Likewise, b knows his own state, so he knows either that he wants beer or that he doesn't want beer. If he knows that he doesn't, he can infer that E is false, similarly to a. If he knows that he does, then he can't infer anything about E, and has to answer "I don't know".

Now, c has all the information he needs to answer either "Yes" or "No". Obviously, if any of the previous answers are "No", then he should also answer "No" or simply not answer. If both earlier answers are "I don't know", he can infer that both a and b want beer. He also knows his own state. If he does not want beer, he will answer "No". If he does, he can infer that E is true, and thus answer "Yes".

2.2 Comments on the analysis

Section 2.1 analyzed a seemingly simple webcomic, but it was found that it required quite a

few formalizations to work it out. However, the above account is incomplete in that I did not spell out all the propositions used and all the inferences drawn. Any complete account of TOM reasoning should spell out all of these but it would require much more space than used here. See Stenning et al. (2008) for a book length attempt at such analyses as part of their larger project of uniting cognitive science and formal logic.

Furthermore, and more relevant to TOM, it required assumptions of how humans reason about how other humans reason. One had to assume that everybody is ideally rational, has self-knowledge about beer desire, and is honest in communicating them. If the situation is more complex involving possibly dishonest humans or non-ideally rational humans, it can quickly get very complex.

For readers more curious about the logical analysis of TOM reasoning, especially concerning second-order TOM reasoning (reasoning about others' beliefs about beliefs or desires), I wish to draw attention to two such cases mentioned by van Bentham (2008). His two cases are:

Case #1:

> You are in a restaurant with your parents, and you have ordered three dishes: Fish, Meat, and Vegetarian. Now a new waiter comes back from the kitchen with three dishes. What will happen?
>
> The children say, quite correctly, that the waiter will ask a question, say: “Who has the Fish?”. Then, they say that he will ask “Who has the Meat?” Then, as you wait, the light starts shining in those little eyes, and a girl shouts: “Sir, now, he will not ask any more!”

The difficulty of this case involves understanding that the waiter has a desire to know three things, and if he is given information about two of them, he can figure out the last one by deductive logic alone. The girl in his example apparently figured this out by assuming the waiter was sufficiently rational to think of the inference.

Case #2:

> Three volunteers were called to the front, and received one coloured card each: red, white, blue. They could not see the others’ cards. When asked, all said they did not know the cards of the others. Then one girl (with the white card) was allowed a

> question; and asked the boy with the blue card if he had the red one. I then asked, before the answer was given, if they now knew the others' cards, and the boy with the blue card raised his hand, to show he did. After he had answered "No" to his card question, I asked again who knew the cards, and now that same boy and the girl both raised their hands …
>
> The explanation is a simple exercise in updating, assuming that the question reflected a genuine uncertainty. But it does involve reasoning about what others do and do not know. And the children did understand why one of them, the girl with the red card, still could not figure out everyone's cards, even though she knew that they now knew.15

The difficulty has to do with keeping track of what each person knows, and reasoning about what they can infer given the information they have. It also involves understanding that when someone asks a question (assuming that person is not asking for tactical reasons) one can infer that he doesn't know the answer. It gets quite complicated handling all this formally. Most humans do it automatically, but it will be a challenge to get AI's to do it well.

I wrote out much of the logical analysis of these two cases as well, and they are quite a bit more complicated than the above and were too long to be discussed here. Both can be found in Kirkegaard (2012).

3 References

- Davis, Wayne, "Implicature", The Stanford Encyclopedia of Philosophy (Spring 2013 Edition), Edward N. Zalta (ed.), http://plato.stanford.edu/archives/spr2013/entries/implicature/.
- Hare, Brian, and Michael Tomasello. "Human-like social skills in dogs?." *Trends in cognitive sciences* 9.9 (2005): 439-444.
- Horn, Laurence R. *A natural history of negation*. Vol. 960. Chicago: University of Chicago Press, 1989.
- Kahneman, Daniel (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
- Kirkegaard, Emil. 2011. "Three logicians walk into a bar: a formal explanation", in

Clear Language, Clear Mind. http://emilkirkegaard.dk/en/?p=2506 Accessed 30th September 2013.

- Kirkegaard, Emil. 2012. “Interesting paper: Logic and Reasoning: do the facts matter? (Johan van Benthem)”, in *Clear Language, Clear Mind.* http://emilkirkegaard.dk/en/?p=3439 Accessed 30th September 2013.
- Jensen, Arthur R. "Bias in mental testing." (1980).
- Malle, B. F. (2002). The relation between language and theory of mind in development and evolution. In T. Givón & B. F. Malle (Eds.), The evolution of language out of pre-language (pp. 265-284). Amsterdam: Benjamins.
- Miller, Carol A. "Developmental relationships between language and theory of mind." *American Journal of Speech-Language Pathology* 15.2 (2006): 142.
- Smullyan, Raymond M., (1986) *Logicians who reason about themselves*, Proceedings of the 1986 conference on Theoretical aspects of reasoning about knowledge, Monterey (CA), Morgan Kaufmann Publishers Inc., San Francisco (CA), pp. 341-352
- Stenning, Keith, and Michiel Van Lambalgen. *Human reasoning and cognitive science*. MIT Press, 2008.
- van Benthem, Johan. "Logic and reasoning: Do the facts matter?." *Studia Logica* 88.1 (2008): 67-84.
- Wikipedia. English Wikipedia. *Doxastic logic.* Accessed 30th of September 2013.