On Building Robot Persons: Response to Zlatev

SELMER BRINGSJORD
Department of Cognitive Science, Department of Computer Science, Rensselaer Polytechnic Institute(RPI), Troy NY 12180, USA; E-mail: selmer@rpi.edu

Abstract. Zlatev offers surprisingly weak reasoning in support of his view that robots with the right kind of developmental histories can have meaning. We ought nonetheless to praise Zlatev for an impressionistic account of how attending to the psychology of human development can help us build robots that *appear* to have intentionality.

Key words: intentionality, meaning, robots, Turning test

Zlatev tells us in the concluding section of his paper that he has ‘presented a rather long, and somewhat loose, argument for a particular answer to’ this pair of questions:

(Q1) Can a robot[1] have meaning (and other properties constitutive of personhood)?
(Q2) If so, how can this be achieved?

If by ‘loose’ Zlatev means ‘invalid,’ he has indeed delivered. If by ‘loose’ he means ‘enthymematic,’ Zlatev has failed. Zlatev does provide an impressionistic account of a better-than-COG (Brooks et al. 1999; Dennett, 1994) method for trying to build humanoid robots that *appear* to have meaning, but the paper is fatally flawed at its core by bad reasoning.

Consider the inferences Zlatev makes from his ‘Wittgenstein-inspired thought experiment,’ which stands at the very core of his case for an affirmative answer to (Q1). Here’s the full description of the gedanken-experiment, in Zlatev’s words:

> Let us consider the following Wittgenstein-inspired thought experiment: a person who has lived a normal life in our community dies and in the autopsy it is discovered that there is some kind of a device instead of a brain in his head. (Zlatev, 2001, p. 160)

Zlatev’s answer to (Q1), given on the basis of this thought-experiment, is a *conditional* ‘Yes:’

(A1) If an artificial autonomous system (a robot) with [(i)] bodily structure similar to ours (in the relevant aspects) has become able to [(ii)] participate in social practices (language games) by [(iii)] undergoing an epigenetic process of cognitive development and socialization, then we may attribute ‘true’ intelligence and meaning to it. (Zlatev, 2001, p. 161)

Minds and Machine **14:** 381–385, 2004.

But how does (A1) follow from the thought-experiment? It's exceedingly hard to say. For starters, the logical structure of (A1) is itself rather murky. For notice that the answer says that 'If ... we *may* attribute...' The 'may' is peculiar. After all, (Q1) didn't read

(Q1′) Are there conditions under which we (some of us?) might say of a robot that it has meaning?

The answer to this question is rather obviously

(Al′) If a robot is TTT-indistinguishable[2] from a human, then some thoroughgoingly rational people might say that that robot has meaning.

(A1′) is demonstrably true, on reasonable construals of it. For example, if God said to us: 'Look, I've created a robotic person who has your sort of mental life,' and then presented us with an astonishingly impressive robot, many of us would say that it has meaning.[3] Likewise, to

(Q1″) Are there conditions under which we (some of us?) would be (epistemologically) entitled to say of a robot that is has meaning?

it would seem obvious that God's proclamation, made as a gleaming TTT-passing robot is presented, implies an affirmative answer. Hence we know that

(A1″) If a robot passes TTT (in certain contexts), then we are epistemically entitled to say that it has meaning.

is true. The question worth tackling is this one:

(Q1*) Can a robot *in fact* have meaning?

To reconstruct (A1) so that, logically speaking, it can be a candidate answer, Zlatev would need to go with something like

(A1*) If a robot has (i) bodily structure similar to ours (in the relevant aspects) and has thereby become able to (ii) participate in social practices (language games) by (iii) undergoing an epigenetic process of cognitive development and socialization, then it *in fact* has true intelligence and meaning.

If Zlatev's reasoning is to have even a fighting chance, either he attempts to show that (A1*) can be derived from this thought-experiment, which would indeed mark a substantive contribution to philosophy of mind, or he is read as attempting to demonstrate that the trivially true (A1′) follows from his hypothetical autopsy. I assume that Zlatev's objective is the former, but let's keep both goals in mind as we proceed.

So what's the argument? Well, its structure is clear. (A1*) is a conditional. Zlatev's strategy is at bottom to assume that the antecedent of (A1*) is true, and to derive the consequent. Zlatev believes the thought-experiment to be powerful because in it, (i), (ii), and (iii) obtain (along with other states of affairs), and the consequent, that the robot has meaning, is by his lights entailed. This is why immediately after presenting the gedanken-experiment he writes:

> Would we on the basis of this decide that we had been fooled all along and that the person was actually a 'brainless' automaton? I believe that the answer is: hardly. (Zlatev, 2001, p. 160)

This argument uses the familiar pair *conditional proof* (from assuming ϕ, and deducing ψ, infer to $\phi \rightarrow \psi$) and universal introduction. The second of these rules of inference allows one to conclude $\forall x\phi$ from the assumption (or fact) that $\phi(a)$ is true, where a is a constant occurring in ϕ. Let's follow Zlatev and denote the robot in the thought-experiment 'r.' Now we can go back to the key properties described in (i)–(iii); for convenience label them $F_{(i)}$, $F_{(ii)}$, $F_{(iii)}$. Then the idea (and working out the full description of this idea is the bulk of Zlatev's paper) is to assume

$$F_{(i)}r \wedge F_{(ii)}r \wedge F_{(iii)}r.$$

Next, we are to derive that r has meaning, which we can denote by 'Mr.' But what, pray tell, confirms

$$\{F_{(i)}r \wedge F_{(ii)}r \wedge F_{(iii)}r\} \vdash Mr?$$

I can't for the life of me figure this out, despite reading Zlatev's paper a number of times, word by word. The consequent of (A1*), represented here by Mr, just doesn't follow from the assumptions that have been made. Not only that, but some have attempted to devise thought-experiments in which $\neg Mr$ holds despite the supposition that a hypothetical robot has the three properties in question (e.g. see Bringsjord, 1999).

Were Zlatev to try to repair things with better reasoning, he would face an uphill battle, to say the least. The reason is his thought-experiment's use of a fiendishly ambiguous term: 'device.' Zlatev probably has in mind 'computer,' but I doubt he would insist on a particular *kind* of computer. So suppose the device in question is a computer, but an abacus, with beads and wire, and a little bird which rapidly manipulates the beads. Under these conditions, I, for one, would be reluctant to affirm any proposition remotely like the view that r has meaning. If the cranium in question were opened up before me, and there was the bird and beads and wire, well, without question I would at the very least suspend my attribution of meaning to the creature that had behaved so normally when alive. But we know that in fact a digital computer, indeed *any* Turing-equivalent computing machine, is very literally an abacus

(Lambek, 1961). This is one of the reasons why I believe the answer to (Q1*) is 'No,' even when the 'can' therein is understood to indicate logical possibility.[4]

Zlatev does deserve praise for explaning in broad strokes how we might exploit knowledge of human development in order to build TTT-passing zombie robots – TTT-passing robots that will *appear* to have intentionality, and that thereby impel some to declare that they do *in fact* have meaning. We owe a special debt to Zlatev for his explanation because it reveals the laughable naivete of the Cog project, the leaders of which hold that sufficiently rich robot development (i.e. development that will impel ascriptions of intentionality to Cog) can be achieved in a laboratory that takes no account of the psychology of human development. In my opinion, by taking account of this part of psychology, Zlatev has offered the best published answer not to (Q2), the second of his driving questions, but rather to

(Q2′) How can we build robots who *appear* to have intentionality?

Notes

[1] Zlatev often says just 'machine' rather than 'computing machine' or 'robot'. But it's evident that he sees himself doing speculative cognitive humanoid robotics. This is why Zlatev says his answers to (Q1) and (Q2) are likely to cause rejoicing in robot fans and alarm in robot foes.
[2] Following Zlatev, I appeal to Harnad's (1991) Total Turing Test, which is passed by a robot if its linguistic *and* sensorimotor powers match those of normal humans. See also Bringsjord (2000a).
[3] Readers may find it worthwhile to read the first known mention of such thought-experiments: Turing (1964).
[4] For arguments supporting this answer see Bringsjord (2000b, 1992).

References

Bringsjord, S. (1992), *What Robots Can and Can't Be*, Dordrecht: Kluwer Academic Publishers.

Bringsjord, S. (1999), 'The Zombie Attack on the Computational Conception of Mind', *Philosophy and Phenomenological Research* 59(1), pp. 41–69.

Bringsjord, S. (2000a), 'Animals, Zombanimals, and the Total Turing Test: The Essence of Artificial Intelligence', *Journal of Logic, Language, and Information* 9, pp. 397–418.

Bringsjord, S. (2000b), 'Clarifying the Logic of Anti-Computationalism: Reply to Hauser', *Minds and Machines* 10, pp. 111–113.

Brooks, R., Breazeal, C., Marjanovic, M., Scassellati, B. and Williamson, M. (1999), The Cog Project: Building a Humanoid Robot, in '*Computat on for Metaphors*, *Analogy*, *and Agents*, Springer-Verlag, Lecture Notes in Computer Science 1562', New York, NY: Springer-Verlag.

Dennett, D. (1994), 'The Practical Requirements for Making a Conscious Robot', *Philosophical Transactions of the Royal Society of London* 349, pp. 133–146.

Harnad, S. (1991), 'Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem', *Minds and Machines* 1(1), pp. 43–54.

Lambek, J. (1961), 'How to Program an Infinite Abacus', *Canadian Mathematical Bulletin* 4, pp. 295–302.

Turing, A. (1964), Computing Machinery and Intelligence, in A.R. Anderson, ed., '*Minds and Machines*', Englewood Cliffs, NJ: Prentice-Hall, pp. 4–30.

Zlatev, J. (2001), 'The Epigenesis of Meaning in Human Beings, and Possibly in Robots', *Minds and Machines* 11, pp. 155–195.