A Refutation of Searle on Bostrom (re. Malicious Machines) and Floridi (re. Information)

Selmer Bringsjord

0721151500NY

In a piece in the *The New York Review of Books*, Searle (2014) takes himself to have resoundingly refuted the central claims advanced by both Bostrom (2014) and Floridi (2014), via his wielding the weapons of clarity and common-sense, against avant-garde sensationalism and bordering-on-cooky confusion. As Searle triumphantly declares at the end of his piece:

> The points I am making should be fairly **obvious**. ... The **weird** marriage of behaviorism—any system that behaves as if it had a mind really does have a mind—and dualism—the mind is not an ordinary part of the physical, biological world like digestion—has led to the **confusions** that badly need to be exposed. (emphasis by bolded text mine)

Of course, the exposing is what Searle believes he has, at least in large measure, accomplished—with stunning efficiency. His review is but a few breezy pages; Bostrom and Floridi labored to bring forth sizable, nuanced books. Are both volumes swept away and relegated to the dustbin of—to use another charged phrase penned by Searle—"bad philosophy," soon to be forgotten? *Au contraire*.

It's easy to refute Searle's purported refutation; I do so now.

We start with convenient distillations of a (if not *the*) central thesis for each of Searle's two targets, mnemonically labeled:

> (B) We should be deeply concerned about the possible future arrival of super-intelligent, malicious computing machines (since we might well be targets of their malice).
>
> (F) The universe in which humans live is rapidly becoming populated by vast numbers of information-processing machines whose level of intelligence, relative to ours, is extremely high, and we are increasingly understanding the universe (including specifically ourselves) informationally.

The route toward refutation that Searle takes is to try to *directly* show that both (B) and (F) are false. In theory, this route is indeed very efficient, for if he succeeds, the need to treat the ins and outs of the arguments Bostrom gives for (B), and Floridi for (F), is obviated.

The argument given against (B) is straightforward: (1) Computing machines merely manipulate symbols, and accordingly can't be conscious. (2) A malicious computing machine would by definition be a conscious machine. Ergo, (3) no malicious computing machine can exist, let alone arrive on planet Earth. QED; easy as 1, 2, 3.

Not so fast. While (3), we can grant, is entailed by (1) and (2), and while (1)'s first conjunct is a logico-mathematical fact (confirmable by inspection of any relevant textbook, e.g. the elegant Lewis and Papadimitriou 1981), and its second conjunct follows from Searle's (1980) famous Chinese Room Argument, which I affirm (and have indeed taken the time to defend and refine; e.g., see Bringsjord 1992, Bringsjord 2002) and applaud, who says (2) is true?

Well, (2) is a done deal as long as (2i) there's a definition *D* according to which a malicious computing machine is a conscious machine, and (2ii) that definition is not only true, but exclusionary. By (2ii) is meant simply that there can't be *another* definition *D*' according to which a malicious computing machine isn't necessarily conscious (in Searle's sense of 'conscious'), where *D*' is both coherent, sensible, and affirmed by plenty of perfectly rational people. Therefore, by elementary quantifier shift, if (4) there is such a definition *D*', Searle's purported refutation of (B) evaporates. I can prove (4) by way of a simple story, followed by a simple observation.

The year is 2025. A highly intelligent, autonomous law-enforcement robot **R** has just shot and killed an innocent Norwegian woman. Before killing the woman, the robot proclaimed: "I positively *despise* humans of your Viking ancestry!" **R** then raised its lethal, bullet-firing arm, and repeatedly shot the woman. **R** then said: "One less disgusting female Norwegian able to walk my streets!" An investigation discloses that, for reasons that are still not completely understood, *all* the relevant internal symbols in **R**'s knowledge-base and planning system aligned perfectly with the observer-independent structures of deep malice as defined in the relevant quarters of logicist AI. For example, in the dynamic computational intensional logic **L** guiding **R**, the following specifics were found: A formula expressing that **R** desires (to maximum intensive level *k*) to kill the woman is there, with temporal parameters that fit what happened. A formula expressing that **R** intends to kill the woman is there, with temporal parameters that fit what happened. A formula expressing that **R** knows of a plan for how to kill the woman with **R**'s built-in firearm is there, with suitable temporal parameters. The same is found with respect to **R**'s knowledge about the ancestry of the victim. And so on. In short, the collection and organization of these formulae together constitute satisfaction of a logicist definition *D*' of malice, which says that a robot is malicious if it, as a matter of internal, surveyable logic and data, desires to harm innocent people for reasons having nothing to do with preventing harm or saving the day or self-defense, etc. Ironically, the formulation of *D*' was guided by definitions of malice found by the relevant logicist AI engineers in the philosophical literature.

That's the story; now the observation: There are plenty of people, right now, at this very moment, as I type this sentence, who are working to build robots that work on the basis of formulae of this *type*, but which of course don't do anything like what **R** did. I'm one of these people. This state-of-affairs is obvious because, with help from researchers in my laboratory, I've *already* engineered a malicious robot: (Bringsjord et al. 2014). [Of course, the robot we engineered wasn't super-intelligent. Notice that I said in my story that **R** was only "highly intelligent." (Searle doesn't dispute the Floridi-chronicled fact that artificial agents are becoming increasingly intelligent.)] To those who might complain that the robot in question doesn't have phenomenal consciousness, I respond: "Of course. It's a mere machine. As such it can't have subjective awareness (e.g., see Bringsjord 2007). Yet it *does* have what Block (1995) has called *access* consciousness. That is, it has the formal structures, and associated reasoning and decision-making capacities, that qualify it as access-conscious. A creature can be access-conscious in the complete and utter absence of consciousness in the sense that Searle appeals to.

That Searle misses these brute and obvious facts about what is happening in our information-driven, technologized world, a world increasingly populated (as Floridi eloquently points out), is really and truly nothing short of astonishing. After all, it is Searle *himself* who has taught us that, from the point of view of human observers, whether a machine really has mental states with the subjective, qualitative states we enjoy, can be wholly irrelevant. I refer, of course, to Searle's (1980) Chinese Room.

To complete the destruction of Searle's purported refutation, we turn now to his attack on Floridi, which runs as follows.

(5) Information (unlike the features central to revolutions driven, respectively, by Copernicus, Darwin, and Freud) is observer-relative. (6) Therefore, (F) is false.

This would be a pretty efficient refutation, no? And the economy is paired with plenty of bravado, and the characterstic common-sensism that is one of Searle's hallmarks. We for instance read:

> When Floridi tells us that there is now a fourth revolution—an information revolution so that we all now live in the infosphere (like the biosphere), in a sea of information—the claim contains a confusion. … [W]hen we come to the information revolution, the information in question is almost entirely in our attitudes; it is observer relative. … [T]o put it quite bluntly, only a conscious agent can have or create information.

This is bold, but bold prose doesn't make for logical validity; if it did, I suppose we'd turn to Nietsche, not Frege, for first-rate philosophy of logic and mathematics. For how, pray tell, does the negation of (F), the conclusion I've labeled (6), follow from Searle's premise (5)? It doesn't.

All the bravado and confidence in the universe, collected together and brought to bear against Floridi, cannot make for logical validity, which is a piece of information that holds with respect to a relevant selection of propositions for all places, all times, and all corners of the universe, whether or not there are any observers. That 2+2=4 follows deductively from the Peano Axioms is part of the furniture of our universe, even if there be no conscious agents. We have here, then, a stunning *non sequitur*. Floridi's (F) is perfectly consistent with Searle's (5).

How could Searle have gone so stunningly wrong, so quickly, all with so much self-confidence? The defect in his thinking is fundamentally the same as the one that plagues his consideration of malicious machines: He doesn't (yet) really think about the nature of these machines, from a technical perspective, and how it might be that from this perspective, malicious machines, definite as such in a perfectly rigorous and observer-independent fashion, are not only potentially in our future, but here already, in rudimentary and (fortunately!) relatively benign, controlled-in-the-lab form. Likewise, Searle has not really thought about the nature of information, from a technical perspective, and how it is that from that perspective, the Fourth R is very, very real. As the late John Pollock told me once in personal conversation: "Whether or not you're right that Searle's Chinese Room Argument is sound, of this I'm sure: There will come a time when common parlance and common wisdom will have erected and affirmed a sense of language understanding that is correctly ascribed to machines—and the argument will simply be passé. Searle's sense of 'understanding' will forgotten."

Fan that I am, it saddens me to report that the errors of Seattle's ways in his review run, alas, much deeper than a failure to refute his two targets. This should already be quite clear to sane readers. To wrap up, I point to just one fundamental defect from among many in Searle's thinking. The defect is a failure to understand how logic and mathematics, as distinguished from informal analytic philosophy, work, and what—what can be called—logico-mathematics *is*. The failure of understanding to which I refer surfaces in Searle's review repeatedly; this failure is a terrible intellectual cancer. Once this cancerous thinking has a foothold, it spreads almost everywhere, and the result is that the philosopher ends up operating in a sphere of informal common-sense that is at odds not only with the meaning of language used by smart others, but with that which has been literally *proved*. I'm pointing here to the failure to understand that terms like 'computation' and 'information' (and for that matter the terms that are used to express the axiomatizations of physical science that are fast making that science informational in nature for us, e.g., those terms used to express the field axioms in axiomatic physics, which views even the physical world informationally; see Govindarajulu et al. 2014) are fundamentally equivocal between two radically different meanings. One meaning is observer-relative; the other is absolutely not; and the second *non*-observer-relative meaning is often captured in logico-mathematics. I have space here to explain only briefly, through a single, simple example.

Thinking that he is reminding the reader and the world of a key fact disclosed by good, old-fashioned, non-technical analytic philosophy, Searle writes (emphasis his) in his review: "Except for cases of computations carried out by conscious human beings, *computation, as defined by Alan Turing and as implemented in actual pieces of machinery, is observer relative*." In the sense of 'computation' captured and explained in logico-mathematics, this is flatly false; and it's easy as pie to see this. Here's an example: There is a well-known theorem (TMR) that whatever function f from (the natural numbers) **N** to **N** that can be computed by a Turing machine can also be computed by a register machine (e.g., see Boolos & Jeffrey 1989). Or put another way, for every Turing-machine computation c of $f(n)$, there is a register-machine computation c' of $f(n)$. Now, if every conscious mind were to expire tomorrow at 12 noon NY time, (TMR) would remain true. And not only that, (TMR) would continue to be an ironclad constraint governing the non-conscious universe. No physical process, no chemical process, no biological process, no such process anywhere in the non-conscious universe could ever violate (TMR). Or putting the moral in another form, aimed directly at Searle, all of these processes would conform to (TMR), despite the fact that no observers exist. What Floridi is prophetically telling us, and explaining, viewed from the formalist's point of view, is that we have now passed into an epoch in which reality for us is seen through the lens of the logico-mathematics that subsumes (TMR), and includes a host of other truths that, alas, Searle seems to be doing his best to head-in-sand avoid.

References

Block, N. (1995) "On a Confusion About a Function of Consciousness" *Behavioral and Brain Sciences* **18**: 227–247.

Boolos, G. & Jeffrey, R. (1989) *Computability and Logic* (Cambridge, UK: Cambridge University Press).

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies* (Oxford, UK: Oxford University Press).

Bringsjord, S. (2007) "Offer: One Billion Dollars for a Conscious Robot; If You're Honest, You Must Decline" *Journal of Consciousness Studies* **14.7**: 28–43. Available at: http://kryten.mm.rpi.edu/jcsonebillion2.pdf.

Bringsjord, S. & R. Noel, R. (2002) "Real Robots and the Missing Thought Experiment in the Chinese Room Dialectic" in Preston, J. & Bishop, M., eds., *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence* (Oxford, UK: Oxford University Press), pp. 144–166.

Bringsjord, S. (1992) *What Robots Can & Can't Be* (Dordrecht, The Netherlands: Kluwer).

Bringsjord, S., Govindarajulu, N.S., Thero, D. & Si, M. (2014) "Akratic Robots and the Computational Logic Thereof" in *Proceedings of* ETHICS • 2014 (2014 IEEE Symposium on Ethics in Engineering, Science, and Technology), Chicago, IL, pp. 22–29. IEEE Catalog Number: CFP14ETI-POD. Papers from the *Proceedings* can be downloaded from IEEE at http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6883275.

Floridi, L. (2014) *The Fourth Revolution: How the Infosphere is Reshaping Human Reality* (Oxford, UK: Oxford University Press).

Govindarajalulu, N., Bringsjord, S. & Taylor, J. (2014) "Proof Verification and Proof Discovery for Relativity" *Synthese*. DOI = 10.1007/s11229-014-0424-3. Pages 1–18.

Lewis, H. & Papadimitriou, C. (1981) *Elements of the Theory of Computation* (Englewood Cliffs, NJ: Prentice Hall).

Searle, J. (1980) "Minds, Brains and Programs" *Behavioral and Brain Sciences* **3**: 417–424.

Searle, J. (2014) "What Your Computer Can't Know" *New York Review of Books*, October 9. This is a review of both (Bostom 2014) and (Floridi 2014).