Published in *Studies in History and Philosophy of Science* (2005) 36: 203–209.
doi:10.1016/j.shpsa.2004.12.013

mark.sprevak@ed.ac.uk

Review of *Views Into the Chinese Room*

Mark Sprevak
University of Edinburgh

7 January 2005

> JOHN PRESTON & MARK BISHOP (EDS.), *Views Into the Chinese Room: New Essays on Searle and Artificial Intelligence*. Oxford: Oxford University Press, 2002, xvi+410 pp., £50.00 (hardback). ISBN 0-19-825057-6.

In contrast to many areas of contemporary philosophy, something like a carnival atmosphere surrounds Searle's Chinese room argument. Not many recent philosophical arguments have exerted such a pull on the popular imagination, or have produced such strong reactions. People from a wide range of fields have expressed their views on the argument. The argument has appeared in *Scientific American*, television shows, newspapers, and popular science books. Preston and Bishop's recent volume of essays reflects this interdisciplinary atmosphere. The volume includes essays from computer science, neuroscience, artificial intelligence, cognitive science, sociology, science studies, physics, mathematics, and philosophy. There are two sides to this interdisciplinary mix. On the one hand, it makes for interesting and fun reading for anyone interested in the Chinese room argument, but on the other, it raises the threat that the Chinese room argument might be left in some kind of interdisciplinary no man's land.

The Chinese room argument (CRA) is an argument against the possibility of Strong artificial intelligence (Strong AI). The thesis of Strong AI is that running a program is sufficient for, or constitutive of, understanding: it is merely in virtue of running a particular program that a system understands. Searle appreciates that understanding is a complex notion, and so he has a particular form of understanding in mind: the understanding of simple stories. It seems intuitively obvious that when I read a simple story in English, I understand that story. One could say that somewhere in

my head there is understanding going on. However, if I read a simple story written in Chinese (a language I do not speak), then there is no understanding going on. What makes the difference between these two cases? The advocate of Strong AI says that the difference lies in the fact that I run a particular program in the case of English stories, and I do not run a particular program in the case of Chinese stories. If the program for understanding Chinese stories were given to me, then I would be able to understand Chinese stories. Similarly, if that program were given to any other sufficiently complex system (for example, a Turing machine), then it too would understand Chinese stories.

Searle's argument against Strong AI is as follows. Imagine a monolingual English speaker inside a room with a rule-book and sheets of paper. The rule-book contains instructions in English on what to do if presented with Chinese symbols. The instructions are of the form: 'If you see Chinese symbol *X* on one sheet of paper and Chinese symbol *Y* on another, then write down Chinese symbol *Z* on a third sheet of paper'. Pieces of paper with Chinese writing are passed into the room and the person inside follows the rules and passes pieces of paper out. Chinese speakers outside the room label the sheets that are passed in 'story' and 'questions', and the sheet that comes out 'answers to questions'. Imagine that the rule-book is as sophisticated as you like, and certainly sophisticated enough that the responses that the person inside the room gives are indistinguishable from those of a native Chinese speaker. Does the person inside the room thereby understand Chinese? Searle claims they do not. No matter how sophisticated the rule-book, or how good the responses, person inside the room will still just be mindlessly shuffling symbols, and failing to understand Chinese.

Searle notes that the Chinese room is a computer, and he identifies the rule-book with the program that it runs. He then reminds us that the thought experiment does not depend on the particular rule-book used. This means that the Chinese room thought experiment demonstrates a failure of the Chinese room computer to understand Chinese *no matter what program it runs*. Since the Chinese room is a universal computer, we can conclude that *no program can be constitutive of understanding*. Hence, Strong AI is refuted. The argument can be rephrased in a slightly different way. Take the best attempt at a program that would be constitutive of understanding, for instance, the best program that the AI research could ever hope to produce. Give that program to the person inside the Chinese room. The program should, in theory, produce Chinese understanding. But we cannot imagine the person inside the room ever understanding Chinese, no matter what program they are given. Hence, no such program that is sufficient for, or constitutive of understanding, can exist.

Searle's argument is clarified on several points that emerge from this anthology.

First, Searle's primary target is functionalism not behaviourism: the primary issue is whether the program that a system runs is sufficient for mentality, not whether intelligent-seeming behaviour alone is sufficient. Second, the subject matter of the argument is metaphysical rather than epistemological: the question is what is constitutive of understanding, not, at least in the first instance, how to test for understanding. Third, Searle uses the term 'understanding' in a way that lends itself to the generation of multiple versions of the Chinese room argument. 'Understanding' could be interpreted as: intentionality, consciousness, a particular mental process, a feeling or quale of understanding, or even mentality in general. Finally, Searle's argument can be thought of as coming in three parts. The first part—the CRA proper—claims that running a program cannot be sufficient for mentality. This claim is compatible with the views of computationalists who claim that an extra ingredient needs to be added to computation in order to produce mentality, e.g. appropriate causal connections to the outside world. The second part of Searle's argument—his response to the Robot reply—denies that adding these extra causal connections would help. The third part of Searle's argument claims that computation, as well as not being a sufficient condition for understanding, is not a necessary condition either.

There are twenty essays in Preston and Bishop's collection. I will focus on just a few of the highlights.

Along with a number of contributors, Ned Block advocates the Systems reply to the CRA. The Systems reply claims that although the man inside the room does not understand Chinese, the system as a whole—the man, plus rule-book, plus pens, plus scraps of paper—does understand Chinese. Searle's response is that the man could in principle memorise the rule-book, perform all the calculations in his head, and be allowed to move around outside the room. The man would then constitute the entire system, but in all other respects the situation would be the same: he would still just be mindlessly shuffling Chinese symbols, and not understanding Chinese. The Systems theorist might respond by saying that although the man may *think* he does not understand Chinese, *in fact* there is a sub-system inside him that does understand Chinese; the man only thinks he cannot understand Chinese because his Chinese sub-system is not hooked up to his English sub-system in the right way. Searle's reply to this is to ask why we should posit such a sub-system. What independent reason do we have for positing a Chinese-understanding sub-personal system? What reason is there apart from the Systems theorist's predisposition to the thesis that the symbol shuffling that the man performs must somehow be sufficient for understanding? Unless the Systems can provide an independent reason, she begs the question against Searle.

Block strengthens the Systems theorist's case by citing multiple personality disorders

as real cases where sub-personal systems can understand. But this would cut no ice with Searle. Searle has nothing against multiple personalities or sub-personal systems that understand. His disagreement is with the reasons for positing such sub-systems. Even if one admits that there are such things as multiple personalities, a Systems theorist would still have to justify, independently of her computationalist assumptions, that what the person in the Chinese room has is a Chinese multiple personality. There seems little that the Systems theorist can do in this respect without again begging the question against Searle. Citing multiple personality strengthens the Systems theorist's case in that it shows that sub-personal systems can understand, but it does not establish that this is the kind of sub-personal system that we are dealing with in the Chinese room.

Jack Copeland gives a different treatment of the Systems reply. Copeland contrasts the Systems reply with what he calls the 'logical reply'. The logical reply claims that there is a logical inconsistency in Searle's argument. Searle's argument starts from the premise that the man inside the room (Copeland calls him 'Clerk') does not, in virtue of his symbol shuffling, understand Chinese. Searle concludes from this that the symbol shuffling carried out by Clerk does not enable the room to understand Chinese. Copeland correctly points out that, as this stands, it is a *non sequitur*. It would be like saying that 'Clerk has no taxable assets in Japan' entails 'The organisation of which Clerk is a part has no taxable assets in Japan'. The logical reply differs from the Systems reply in that the logical reply does not have a positive component: it consists only in pointing out an error in Searle's argument. The logical reply is neutral on the issue of whether the room as a whole understands Chinese.

The logical reply can be transposed to the sub-system argument of the Systems reply. Copeland suggests that Searle blocks this move by implicitly assuming what Copeland calls the 'Part-Of principle': 'If Clerk does not understand the Chinese story, then no part of Clerk understands the Chinese story'. The Part-Of principle does block the logical reply, but I am not sure that Searle accepts such a principle. As Copeland admits, Searle has no objection to sub-personal systems that understand. What Searle objects to is that there is sufficient evidence, given the setup of the Chinese room, for positing such a sub-personal system. This could be called the Part-Of* principle: 'If Clerk does not understand the Chinese story then, given the setup of the Chinese room, we have no good reason for thinking that any part of Clerk can understand the Chinese story'. Copeland criticises Searle for not producing an independent justification of the Part-Of principle, and he might feel similarly unhappy about the Part-Of* principle. However, this is not an easy area in which to score points against Searle. It is true that there is little Searle can say to independently justify the Part-Of* principle, but it is equally true that there is little that a critic can say to show that the Part-Of* principle is false. The argument between Searle and his critics seems to reach an impasse at this point.

Georges Rey considers a different strategy for criticising Searle. Rey asks us to imagine a robot that can move about the world and that is functionally as well as behaviourally identical to a human. Rey notes that if we were to encounter such a robot then, from an explanatory point of view, we would subsume it under the same laws and counterfactuals as we do for a normal Chinese speaker. Such a robot would appear to add by carrying just as a Chinese speaker does, it would appear to maximise its expected utilities just as a Chinese speaker does, and so on. So just as we infer from the fact that a Chinese speaker fits these regularities and counterfactuals that she genuinely understands, so too we should infer that the robot genuinely understands. It seems little more than biological chauvinism to grant mentality to the human while denying it to the robot.

The disagreement between Rey and Searle lies at a fundamental level. Rey's argument depends on a third person point of view of mentality, while Searle's argument depends on a first person point of view. Rey's argument is that, from an external point of view, we are equally justified in attributing mentality to the robot and the Chinese person. This works as an argument for robot mentality only if the third person point of view has priority in this instance—in other words, if evidence from a third person point of view is decisive in this judgement of mentality. This is exactly what Searle rejects, and the CRA is an example of how, for him, third person and first person judgements can come apart. If one were to assume that the third person point of view has priority from the start, then the Chinese room argument would not even be given a chance. Understandably, Searle would not be moved by such an argument. Unless the third person critic can find some common third person ground with Searle on which to criticise him, she risks begging the question against him.

Larry Hauser works hard to find such third person common ground in his essay. Hauser's conclusion is that the sincere report of the man inside the Chinese room that he does not understand Chinese should be overruled by third person evidence to the contrary. The first step in Hauser's argument is to deny Searle's distinction between as-if Chinese intentionality (which can be attributed to the room) and intrinsic Chinese intentionality (which only real Chinese people have). Hauser claims that standard linguistic tests for ambiguity yield no evidence of ambiguity between our concept of as-if intentionality and our concept of intrinsic intentionality, and therefore we should not assume that there is a difference. If this is the case, then there is only one type of intentionality at issue in the CRA. This means that one can overrule the claims of the man inside the room that he does not have Chinese intentionality with the overwhelming third person evidence that he has. Hauser's roundabout approach goes some way towards addressing third person worries mentioned above, but it does not go far enough. Searle is likely to protest at the weight Hauser gives to particular linguistic tests for ambiguity. Intuitions about

ambiguity are controversial and Searle is unlikely to allow the results of such tests to counterbalance what is in his view overwhelming first person evidence to the contrary.

Instead of assuming the priority of the third person point of view, a critic of Searle might try to go on the offensive and undermine the priority of Searle's first person point of view. In her essay, Diane Proudfoot attempts to do just this. Proudfoot begins by describing similarities between the CRA and Wittgenstein's remarks on machines and understanding. She then goes on to claim that although Wittgenstein would agree with Searle's anti-AI conclusion he would not accept Searle's argument for that conclusion. In Proudfoot's opinion, Wittgenstein would agree with the anti-AI conclusion because, for him, understanding cannot be a process, and so *a fortiori* it cannot be a process of symbol manipulation. According to Proudfoot, Wittgenstein would reject Searle's argument because he would object to the CRA's reliance on the first person perspective. Wittgenstein held that the criteria on whether a given individual understands are often external to that individual, and according to Proudfoot, this means Searle's first person point of view cannot be assumed to have priority.

I agree that Wittgenstein would reject Strong AI because of his belief that that understanding cannot be process. I also agree that Wittgenstein emphasised the importance of externalist criteria when we decide whether someone understands. But I think his Wittgenstein's point was not, as Proudfoot seems to suggest, a point about the externalist nature of understanding or mental content. Wittgenstein was not saying that there is one thing or property, understanding, that has externalist as well as internalist criteria. Rather, I think that Wittgenstein's concern was to point out that we have many different uses for the word 'understanding', some of which involve externalist criteria and some of which do not (in some situations we do regard the person to be authoritative on whether she understands). I think that it would be unWittgensteinian to assume that all of these different uses of 'understanding' name the same thing or property, or name at all. In my view, Wittgenstein was not an externalist about content and understanding. Therefore, I find it hard to see how Wittgenstein can be used to show that Searle's use of the first person point of view in this instance must be wrong.

This is a small sample of the issues that appear in the collection. Other highlights include discussions by Proudfoot and Wheeler on how embedded cognition and dynamical systems approaches relate to the CRA; Bringsjord and Noel's attempt to provide a thought experiment to defend Searle against the combination of the Systems and Robot replies; Harnad's clear discussion of the strengths and weaknesses of the CRA; Preston's historical overview of AI and the CRA; Haugeland's discussion of program syntax and semantics; and Searle's own contribution. There are many

other issues worth discussing, but I wish to return to the worry mentioned in the introductory paragraph: that an interdisciplinary approach might leave the Chinese room argument in a no man's land.

Preston and Bishop do not explicitly justify the interdisciplinary approach of their book, but there is much that could be said about the merits of an interdisciplinary approach. However, there are problems as well as benefits associated with an interdisciplinary study. One potential problem is that the reader may be left confused about the exact role of the CRA. Why does the CRA merit an interdisciplinary study? The answer is not clear. Compounding this difficulty is the widely held belief that the CRA is either wrong, irrelevant, or in some cases, pernicious. Preston mentions this worry in his introduction:

> In preparing this volume, the editors became more aware than ever of a sort of consensus among cognitive scientists to the effect that the CRA is, and has been shown to be, bankrupt … Some prominent philosophers of mind declined to contribute on the grounds that the project would give further exposure to a woefully flawed piece of philosophizing. Even some of those who have contributed to the volume think of the CRA not just as flawed, but as pernicious and wholly undeserving of its fame. (Preston and Bishop 2002, pp. 46–47)

Preston's remarks are accurate, but depressing. They reflect a widespread, and I think erroneous, estimate of the power and adaptability of Searle's argument. Sadly, one of the major challenges that the Chinese room argument faces today is to continue to justify its own existence. My main concern with Preston and Bishop's anthology is that it does not answer this worry. It is hard to know exactly how to deal with this problem. My personal view is that the answer is unlikely to be an interdisciplinary study. The CRA needs to survive by consolidating, not expanding. I think that what is needed is for a case to be made for the CRA's importance to current philosophical issues. Once this rationale in place, then an interdisciplinary study can follow. The first step is currently what is lacking.

I would not hesitate to recommend Preston and Bishop's collection: it has a number of very good essays, its mix of styles makes for interesting reading, and it has an excellent bibliography. My only worry is that its interdisciplinary nature, while immediately stimulating, risks reinforcing the CRA's homelessness over the longer term. The CRA is fun, but we should take care that after the whirl of the carnival has passed, people still care about the CRA, and can justify why they do so.

Acknowledgements

I would like to thank Peter Lipton and an anonymous referee for comments on a previous draft of this review.

References

Preston, J., and M. Bishop, eds. 2002. *Views into the Chinese Room.* Oxford: Oxford University Press.