Preston, John, & Bishop, Mark (2002), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence* (Oxford: Oxford University Press), xvi + 410 pp., ISBN 0-19-825057-6.

Reviewed by:

William J. Rapaport

Department of Computer Science and Engineering, Department of Philosophy, and Center for Cognitive Science, State University of New York at Buffalo, Buffalo, NY 14260-2000; `rapaport@cse.buffalo.edu`, `http://www.cse.buffalo.edu/∼rapaport`

This anthology's 20 new articles and bibliography attest to continued interest in Searle's (1980) Chinese Room Argument. Preston's excellent "Introduction" ties the history and nature of cognitive science, computation, AI, and the CRA to relevant chapters.

Searle ("Twenty-One Years in the Chinese Room") says, "purely . . . syntactical processes of the implemented computer program could not by themselves . . . *guarantee* . . . semantic content . . . essential to human cognition" (51). "Semantic content" appears to be mind-external entities "attached" (53) to the program's symbols. But the program's implementation must accept these entities as input (suitably transduced), so the program in execution, accepting and processing this input, *would* provide the required content. The transduced input would then be *internal* representatives of the external content and would be related to the symbols of the formal, syntactic program in ways that play the same roles as the "attachment" relationships between the *external* contents and the symbols (Rapaport 2000). The "semantic content" could then just be those mind-*internal* relationships, thus syntactic. Searle disagrees: The CRA "rests on two . . . logical truths . . . [S]yntax is not semantics. . . . [S]imulation is not duplication" (52). However, denying these isn't *logically* inconsistent: Semantic correspondence between domains *can* arise from symbol manipulation, as just suggested, and simulation *can* be duplication, at least in the special cases of cognition and information: *Pace* Searle, simulated information *is* real information (consider a photocopy of a book), because an item's informational content lies in its abstract structure; structure-preserving simulations thereby contain the same information (Rapaport 1988 and in-press).

Block ("Searle's Arguments against Cognitive Science") complains that mechanically executing a natural-language-understanding program doesn't *feel* like language understanding: "[W]hen you seem to Chinese speakers to be ... discours[ing] ... in Chinese, all you are aware of doing is thinking about what noises the program tells you to make next, given the noises you hear and what you've written on your mental scratch pad" (72). This nicely describes the situation of novice second-language speakers: consciously and laboriously computing what their interlocutor says and how to respond. Does this non-native speaker understand the language? The non-native speaker (or Searle-in-the-room) might say "no", while the native speaker says "yes". The native speaker's judgment should prevail (Turing 1950, Rapaport 2000), because ...

... as Hauser ("Nixin' Goes to China") notes, one can understand without *feeling* that one understands. The CRA is based on a "dubious" principle that "first-person disavowals of understanding" are "epistemically privileged" (127–128)—dubious because it does not follow from "Searle's seeming to himself not to understand" (128) that Searle does not really understand.

For Winograd ("Understanding, Orientations, and Objectivity"), there is *no* "answer to ... whether the computer (or the Chinese Room) 'understands' language" (80); language "isn't prepared" for this: "We have clear intuitions that ... pencil sharpeners ... don't understand, and that human[s] ... do. But the computer is a mix-and-match" (82). Winograd notes, correctly, that there are many different senses of 'understand'. But, fixing a sense of 'understand', we can ask of a computer or the CR whether it understands *in that sense*. Our judgment should be based on the same criteria in both cases (Turing 1950). You or the computer understand a newspaper editorial in one sense if you can answer standard reading-comprehension questions, in another if you "get" the political undertones. The computer's understanding should be no different.

Simon & Eisenstadt ("A Chinese Room that Understands") claim to provide such precise tests for NLU, viz., translation ability, NL question-answering, and "similar tasks" (95). But their claim that "*a* computer ... has been programmed ... to understand" NL (95; my emphasis) underwhelms: They present *three* programs, *each* of which implements a *partial* theory of NLU (101). Nevertheless, this is progress.

Where Simon & Eisenstadt (and Aleksander, below) use real computer programs to *contradict* the CRA, Bringsjord & Noel ("Real Robots and the Missing Thought-Experiment in the Chinese Room Dialectic") claim that "real robots ... *strengthen*" it (145). "[Z]ombanimals" (145) are real robots "displaying our external behavior" without consciousness" (157f; cf. 146). We are supposed to conclude that Searle-in-the-room equipped with “the entire system” appears to see things with understanding, but “clearly” (164) does not. ‘Clearly’ needs clarification: Perhaps the equipment needed to *appear* to see with understanding would be no different from *really* seeing with understanding.

Winograd’s “mix and match” status is explored in Adam’s and Warwick’s chapters. For Warwick (“Alien Encounters”), “simple human biases” underlie the CRA, in part because it’s possible to have a conscious machine whose consciousness does not arise from a computer program—witness “insect-like” robots. But Warwick offers no evidence that these robots are *not* computational, nor that what they do is not comput*able* (see Harnad, below).

Adam (“Cyborgs in the Chinese Room”), echoing Warwick, advocates blurring the human-machine boundary (322) separating the “profane”/“unclean”/“them” from the “holy”/“clean”/“us” (Douglas 1966). AI seen “as outside the human body” is “profane and unholy” (326). Boundaries produce marginal members, viewed as dangerous or powerful. “[A]n alternative reading of machines in our society … include[s] them as marginal members … offer[ing] an explanation of why they are potentially threatening … ” (326). Actor-network theory (Latour 1987) is like the Systems Reply, blurring this boundary by considering “the process of creating scientific and technical knowledge in terms of a network of actors … where power is located throughout the network rather than in the hands of individuals” (331). On the other hand, “cyborg feminism” (Haraway 1991) blurs the boundary by definition, since a “cyborg” is a “fabricated hybrid of machine and organism” (334), appearing to be a version of the Robot Reply.

For Proudfoot (“Wittgenstein’s Anticipation of the Chinese Room”), Searle-in-the-room’s Chinese utterances are not speech acts (167): “asking in Chinese ‘Are you in pain?’ when the speaker does not know Chinese” is “a paradigm example of talking ‘without thinking’ ” (167, citing Wittgenstein 1989). Surely, a native Chinese speaker *could* sincerely ask Searle-in-the-room if he is in pain. Could Searle-in-the-room sincerely *answer*? If sincerely asked in English, *I* sincerely answer by knowing what ‘pain’ means, introspecting to see if I’m in pain, then answering. But who or what introspects in the CR? To whom or what should the interrogator’s uses of ‘you’ refer (and Searle-in-the-room’s uses of ‘I’)? If “you” is the system (Searle plus Chinese instruction book), where would the pain (if any) be located, and how would it be sensed?

Rey (“Searle’s Misunderstandings of Functionalism and Strong AI”) may have a key to handling Proudfoot’s problem: The instruction book must “relate Chinese characters not only to one another, but also to the inputs and outputs of the *other* programs [that the “Computational-Representation Theory of Thought” (CRTT)] posits to account for the *other* mental processes and propositional attitudes of a normal Chinese speaker” (208). However, Rey also says, “There’s no reason whatever to suppose that the functional states of a *pain program memorizer* are the same as those of *someone actually in pain*" (214). Rey's chapter, rich in ideas and bearing detailed study, claims that the CRA is irrelevant to CRTT (203), because the Turing Test is behavioristic, concerned only with external input and output, and not "committed to ... a Conversation Manual Theory of Language" (207f), since a *lot* more is needed than a "Berlitz Conversation manual" (208) (cf. Rapaport 1988). *Several* instruction books are needed, corresponding to interacting modules for various cognitive abilities, probably being executed in parallel, hence by more than one inhabitant of the Room (see Taylor, below). In this situation, possibly what's happening inside the room *would* be functionally equivalent to normal Chinese understanding. To the extent that the TT doesn't care about this, too bad for the TT and the CRA.

Harnad ("Minds, Machines, and Searle 2") defines "computationalism" via three "tenets": "(1) Mental states are just implementations of (the right) computer program(s) [which must] be *executed* in the form of a dynamical system" (297). But rather than "Computationalism [being] the theory that cognition is comput*ation*" (297, my italics), it should be the theory that cognition is comput*able* (Rapaport 1998). While mental states are implementations of (the right) computer *processes*, *human* cognition could result from non-computational brain processes—as long as the behavior is *also* characterizable computationally. Perhaps this is *part* of tenet (2): "Computational states are implementation-independent" (297), implying that "if all physical implementations of ... [a] computational system are ... equivalent, then when any one of them has (or lacks) a given computational property [including "being a mental state"], it follows that they all do" (298). But equivalent in input-output behavior, algorithm, data structures? Two such implementations might be *weakly* equivalent if they have (only) the same input-output behavior; degrees of *strong* equivalence might depend on how alike the intervening computer programs were in terms of algorithms, subroutines, data structures, complexity, etc. Tenet (3) is that TT-indistinguishability is the strongest empirical test for the presence of mental states (298). Harnad, echoing Rey, admits that this is input-output, i.e., weak, equivalence (299).

Taylor ("Do Virtual Actions Avoid the Chinese Room?") presents the CRA via slaves carrying out the Chinese NLU algorithm, suggesting an interesting variation on the Systems Reply: Here, no single person can claim, as Searle-in-the-room does, that he (or she) doesn't understand Chinese, yet Chinese is being understood. Thus, either *the entire system* (not any of its components) understands Chinese, or *nothing* does the understanding, despite understanding *happening*. Taylor meets Searle's challenge with "a neurally based ... semantics" (270). If Taylor means that one *neural representation* (of a word) is correlated with another *neural representation* (of an object), I approve. Unfortunately, he postulates that this is the site of Chomsky's (1965) deep structures, a theory no longer defended.

Bishop ("Dancing with Pixies") offers a weaker version of "Putnam [1988]'s claim that, 'every open system implements every Finite State Automaton (FSA)', and hence that psychological states of the brain cannot be functional states of a computer" (361): "over a finite time window, every open system implements the trace of a particular FSA. ... lead[ing] to panpsychism" and, by a *reductio*, "a suitably programmed computer *qua* performing computation can [n]ever instantiate genuine phenomenal states" (361). His argument is odd: For any Discrete State Machine (one capable of different output behavior depending on its input), there will be several other machines, each with *fixed* input, such that each of these machines' output matches the DSM's output for the appropriate input. Then, for any *cognitive* DSM, "we can generate a corresponding state transition sequence using any open physical system" (368). But suppose that the cognitive DSM is "collapsed" into several of these state transition sequences (presumably one per possible input). Choose one. Find an (arbitrary) "open physical system" that has that same state transition sequence. It doesn't follow that that system is cognitive *just* because it does *part* of what the cognitive DSM does.

Haugeland ("Syntax, Semantics, Physics") explores the Systems Reply: Searle "asks himself what it would be like if he were *part of* a mind that worked according to the principles that strong AI says all minds work on—in particular, what it would be like if he were the central processing unit" (379). But "neither the question nor the answer [viz., that the CPU does *not* understand Chinese] is very interesting" (379), since the CRA commits both part-whole and equivocation fallacies (382).

Coulter & Sharrock ("The Hinterland of the Chinese Room") assert: "If computation requires intelligence, and ... can be done on machines, then, [Turing] ... thought, since machines can do computation they must possess intelligence" (184). No: Turing 1936 argues that computation does *not* require intelligence; Turing 1950 argues for the *converse*.

Penrose ("Consciousness, Computation, and the Chinese Room") says that "there must be *non-computational* physical actions underlying the brain processes that control our mathematical thought processes ... [and] that underlie our *awareness*" (236–237) because of his infamous Gödelian argument. But a human or a computer could use *two* formal systems, a proof-theoretic one and a corresponding model-theoretic one, both of which are syntactic (i.e., symbol-manipulation) systems, such that the former cannot prove some well-formed formula, while the latter determines that it is true.

Finally, others discuss variations on "computation": Wheeler's "Change in the Rules" concerns dynamical systems, Copeland's "The Chinese Room from a Logical Point of View" discusses "hypercomputation", and Aleksander's "Neural

Depictions of ‘World’ and ‘Self’ ” considers “neurocomputing”.

Despite Searle’s sentiment that he’s finished with the CRA, “there is (still) little agreement about exactly how the argument goes wrong, or about what should be the exact response on behalf of computational cognitive science and Strong AI” (Preston, p. 47). The CRA is an easy-to-understand and engaging argument around which a host of important philosophical issues can be approached. This book is a good place to explore them.

References

Chomsky, Noam (1965), *Aspects of the Theory of Syntax* (Cambridge, MA: MIT Press).

Douglas, M. (1966), *Purity and Danger: An Analysis of the Concepts of Pollution and Taboo* (London: Ark).

Haraway, Donna (1991), “A Cyborg Manifesto: Science, Technology and Socialist-Feminism in the Late Twentieth Century”, *Socialist Review* 80: 65–107.

Rapaport, William J. (1988), “Syntactic Semantics: Foundations of Computational Natural-Language Understanding”, in James H. Fetzer (ed.), *Aspects of Artificial Intelligence* (Dordrecht, Holland: Kluwer Academic Publishers): 81–131.

Rapaport, William J. (1998), “How Minds Can Be Computational Systems”, *Journal of Experimental and Theoretical Artificial Intelligence* 10: 403–419.

Rapaport, William J. (2000), “How to Pass a Turing Test: Syntactic Semantics, Natural-Language Understanding, and First-Person Cognition”, *Journal of Logic, Language, and Information* 9(4): 467–490.

Rapaport, William J. (in press), “Implementation Is Semantic Interpretation: Further Thoughts”, *Journal of Experimental and Theoretical Artificial Intelligence*.

Searle, John R. (1980), “Minds, Brains, and Programs”, *Behavioral and Brain Sciences* 3: 417–457.

Turing, Alan M. (1936), “On Computable Numbers, with an Application to the Entscheidungsproblem”, *Proceedings of the London Mathematical Society*, Ser. 2, Vol. 42: 230–265.

Turing, Alan M. (1950), “Computing Machinery and Intelligence”, *Mind* 59: 433–460.

Wittgenstein, Ludwig (1989), *Wittgenstein’s Lectures on Philosophical Psychology 1946-1947*, ed. P.T. Geach (Chicago: University of Chicago Press).