UNGROUNDED SEMANTICS: 
SEARLE'S CHINESE ROOM THOUGHT EXPERIMENT, 
THE FAILURE OF META- 
AND SUBSYSTEMIC UNDERSTANDING, AND 
SOME THOUGHTS ABOUT THOUGHT-EXPERIMENTS

CHRISTIAN BEENFELDT

University College, Oxford
The Danish National Research Foundation: Center for Subjectivity Research
University of Copenhagen

1. Syntax and Semantics in the Chinese Room Argument

With the 2002 anthology Views Into the Chinese Room published by Oxford University Press, John Searle’s famous Chinese Room Argument was subject to yet another round of debate – some 22 years since its original introduction in an 1980 issue of the Behavioral and Brain Sciences. From its very first appearance the argument clearly touched a nerve – as evidenced by the 27 peer-review commentaries that Searle’s paper immediately elicited – and in subsequent years, hundreds of articles discussing the argument have been written by philosophers, psychologists, cognitive scientists and Artificial Intelligence researchers alike, with evaluations of the argument spanning the entire spectrum from the very critical to the highly laudatory. Preston goes as far as to characterize the argument as “contemporary philosophy’s best-known argument” and Dietrich states that “AI would have no future if Searle is correct.” Although both are overselling the point here, there is widespread agreement about the substantial importance of Searle’s argument to the issue of machine thinking. As Teng puts it:

Searle’s... Chinese room argument has been one of the most celebrated philosophical arguments in the contemporary philosophy of mind. It is simple and elegant, and is the best-known and most-cited philosophical argument against Strong AI.

Searle’s argument runs as follows. Imagine yourself locked in a room with several baskets full of Chinese symbols and assume that you do not understand a word of Chinese. You are given a rulebook in English for manipulating Chinese symbols strictly in terms of their formal characteristics and intersymbolic
interrelationship, not their meaning. Now some Chinese symbols are passed into the room and you are instructed to send Chinese symbols out of the room, all in accordance with the rulebook specifications, where a rule might say, for example, "Take a squiggle-squiggle sign out of the basket number one and put it next to a squoggle-squoggle sign from basket number two." What really goes on, you later learn, is that a group of computer programmers have written a program to simulate the understanding of Chinese, enabling the entity executing it to answer questions in that language by matching the input against its memory and producing a suitable prefabricated Chinese response as the output. Unbeknown to you, people outside the room refer to the symbols send into the room as "questions" and the symbols you pass back out of the room as "answers to questions" and since the rules are so well written, the symbols so well arranged, and your symbol manipulations conducted with such a degree of expeditious adroitness, the room as a whole achieves indistinguishability with a native speaker of Chinese. If this is all the case, you – plus the room, the rules and the baskets of symbols – will in principle be able to pass the Turing Test, and the room (with all its viscera, including you) will accordingly be regarded as an intelligent speaker of Chinese. Yet, the point the story is that while to external observers it seems as if there is someone in the room who understands Chinese, you have no such understanding – and, worse still, there is apparently no way for you ever to learn Chinese by the method of simply manipulating such formal symbols. If, accordingly, you can complete a sequence of discretely specified symbol manipulation routines while being utterly incognizant of their meaning, and thus can pass a Turing Test, then so can a digital computer; it is, indeed, the very ability to perform calculating operations on any purely formal computational content that is the very source of power for such a device.

To carve the point about the absence of meaning inside the Chinese Room into high relief, contrast the process of manipulating uninterpretable symbols according to strictly formal characteristics, with the process of answering questions in English, where one is cognizant not only of the formal characteristics of the symbols but also of their meaning. For example, compare replying 000100101011 to the question 00111101101010001, with answering, say, "toast" to the question "what did you eat for breakfast?" The intuitive thrust of the Chinese Room Argument thus derives from the fact that the Chinese symbols appear to the symbol manipulator as mere syntax, consisting exclusively of formal characteristics and patterns of formation, while English symbols also