BEHAVIORAL AND BRAIN SCIENCES (2004) **27,** 153–168
Printed in the United States of America

Continuing Commentary

Commentary on **John Searle (1980). Minds, Brains, and Program. BBS 3(3):417–57.**

Abstract of the original article: This article can be viewed as an attempt to explore the consequences of two propositions. (1) Intentionality in human beings (and animals) is a product of causal features of the brain. I assume this is an empirical fact about the actual causal relations between mental processes and brains. It says simply that certain brain processes are sufficient for intentionality. (2) Instantiating a computer program is never by itself a sufficient condition of intentionality. The main argument of this article is directed at establishing this claim. The form of the argument is to show how a human agent could instantiate the program and still not have the relevant intentionality. These two propositions have the following consequences: (3) The explanation of how the brain produces intentionality cannot be that it does so by instantiating a computer program. This is a strict logical consequence of 1 and 2. (4) Any mechanism capable of producing intentionality must have causal powers equal to those of the brain. This is meant to be a trivial consequence of 1. (5) Any attempt literally to create intentionality artificially (strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain. This follows from 2 and 4.

"Could a machine think?" On the argument advanced here, only a machine could think, and only very special kinds of machines, namely, brains and machines with internal causal powers equivalent to those of brains. And that is why strong AI has little to tell us about thinking, since it is not about machines but about programs, and no program by itself is sufficient for thinking.

The Chinese room is a trick

Peter Kugel

Computer Science Department, Boston College, Chestnut Hill, MA 02467-3808. **Kugel@bc.edu** **http://www.cs.bc.edu/~kugel/**

Abstract: To convince us that computers cannot have mental states, Searle (1980) imagines a "Chinese room" that simulates a computer that "speaks" Chinese and asks us to find the understanding in the room. It's a trick. There is no understanding in the room, not because computers can't have it, but because the room's computer-simulation is defective. Fix it and understanding appears. Abracadabra!

In his target article "Minds, Brains, and Programs," Searle (1980) argues that, although computers can seem to have mental states, they cannot really have them. To support his claim, he asks us to imagine a "Chinese room" that (1) simulates what computers can do (2) to produce the appearance of understanding Chinese (3) without having anything that corresponds to "understanding" inside.

Most of those who have argued against Searle – and there have been many – have accepted (1) and (2) and have tried to find "understanding" in the room. That's a mistake because Searle is right. It's not there.

Understanding is not missing because computers can't have it. It's missing because claim (1) – that the room can do everything computers can – is false. The room's computer-imitation is so poor that claim (2) – that the room can do a good job of faking fluent Chinese – is also false.

To see how limited its (apparent) understanding of Chinese is, consider the following dialogue, translated into English for my (and, presumably, most readers') convenience:

> Me: "From here on in I'm going to use the word 'bad' to mean 'good' as it does in some contemporary American slang. Got it?"
> Room: "Yes."
> Me: "Would you say that an A was a bad grade?"
> Room : "No." (Gotcha!)

The reason the room can't handle this sort of thing is that it cannot write anything its user (Searle) can read. According to Searle, it can only write Chinese characters – which Searle cannot read. Which is why it cannot remember things like my "bad" news.

If we allowed the script to change the script (as a computer can change its program), it could change the room's behavior in response to events. That would make the script a lot more complicated, but it would make intentionality possible. And it is intentionality that, according to Searle (1980) and Brentano (1874/1973), distinguishes mental states from physical ones.

According to Searle (1980), internal states have intentionality if they are "directed at or about objects and states of affairs in the world." Let me suggest that what this means is that internal states can change appropriately when what they are "directed at" changes. For example, my thoughts about the Chinese room have an intentionality that is lacking in the Chinese room's "thoughts" about me because my thoughts about the room can change when I learn that it's painted green. The room's thoughts about me lack intentionality because they cannot change when I tell the room that I'm (temporarily) using "bad" differently.

Other mental states have intentionality for similar reasons. For example, what gives my belief that "All swans are white" intentionality is that, after I see a few black swans, my belief can change appropriately, perhaps to "All swans are black or white."

Not all changes produced by experience are sufficiently complex and flexible to count toward intentionality. A supermarket scanner that changes its internal state in response to the UPC code on a bag of cookies lacks intentionality, whereas a Chinese child that changes its internal state in response to the Chinese translation of "I brought home a bag of cookies" has intentionality, as any parent knows.

The Chinese room and the scanner lack intentionality because they only have what I have called "fake intelligence" (Kugel 2002) – the ability to apply the rules (programs, scripts) they have been given. In contrast, a child has intentionality, or "genuine intelligence," because it can adjust, or even build new rules, on the basis of its experiences. And that takes a kind of memory that the Chinese room lacks.

It is not easy to spell out what kinds of changes in response to experiences demonstrate intentionality other than to say that they have to have a certain richness. If philosophers could clarify that

 0140-525X/04 $12.50

(and I believe they can), and if computer scientists could implement programs that can make sufficiently "rich" changes as the result of what they "experience," I would probably call some of those programs' states "mental."

Searle might not. He might still object that the Chinese room, changing its programs in response to its experiences, lacked intentionality because Searle, inside the room, lacked it. That would not bother me because I believe that the intentionality of a human mind does not percolate down to the individual neurons and that, likewise, the intentionality of a computer need not percolate down to its individual components.

Searle might also object that the resulting "understanding" would not feel, to the computer, the way understanding does to him. Since I can only guess how "understanding" feels to Searle, I do not feel competent to comment on this. But, if using the same term for both human and machine states bothers Searle, I would be willing to limit my use of the term "mental states" to refer to what human beings have and to call what computers have "intentional states."

I agree with Searle that machines will have to have something like intentional states before they can become really intelligent. The ability to remember what happens, and to change the way you think in response, is crucial to both intelligence and understanding. You understand this commentary to the degree that it changes what you can do – argue against it, discuss it at cocktail parties, apply its suggestions, and the like.

The English word "mind" is both a noun and a verb. To mind the store is to pay attention to it and change what you are doing in response to what happens in it. I believe that mental states are states that support such minding, and I agree with Searle that programs that lack them cannot be intelligent.

What I do not believe is that such states must be biological.

References

Brentano, F. (1874/1973) *Psychologie vom empirischen Standpunkt,* ed. O. Kraus. Duncker and Humbolt. (1973 English edition: *Psychology from an empirical standpoint,* trans. A. C. Rancurello, D. B. Terrell & L. L. McAlister; ed. L. L. McAlister. Routledge and Kegan Paul/Humanities Press. [PK]

Kugel, P. (2002) Computing machines can't be intelligent (. . . and Turing said so). *Minds and Machines* 12(4):563–79. [PK]

Searle, J. R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3):417–57. [PK]

John Searle has declined to respond to the above continuing commentary.

Commentary on **Linda Mealey (1995). The sociobiology of sociopathy: An integrated evolutionary model. BBS 18(3):523–99.**

Abstract of the original article: Sociopaths are "outstanding" members of society in two senses: politically, they draw our attention because of the inordinate amount of crime they commit, and psychologically, they hold our fascination because most of us cannot fathom the cold, detached way they repeatedly harm and manipulate others. Proximate explanations from behavior genetics, child development, personality theory, learning theory, and social psychology describe a complex interaction of genetic and physiological risk factors with demographic and micro environmental variables that predispose a portion of the population to chronic antisocial behavior. More recent, evolutionary and game theoretic models have tried to present an ultimate explanation of sociopathy as the expression of a frequency-dependent life strategy which is selected, in dynamic equilibrium, in response to certain varying environmental circumstances. This paper tries to integrate the proximate, developmental models with the ultimate, evolutionary ones, suggesting that two developmentaly different etiologies of sociopathy emerge from two different evolutionary mechanisms. Social strategies for minimizing the incidence of sociopathic behavior in modern society should consider the two different etiologies and the factors that contribute to them.

The sociobiology of sociopathy: An alternative hypothesis

Wim E. Crusio

Brudnick Neuropsychiatric Research Institute, Department of Psychiatry, University of Massachusetts Medical School, Worcester, MA 01604.
wim.crusio@umassmed.edu

Abstract: Mealey argued that sociopathy is an evolutionary stable strategy subject to frequency-dependent selection – high levels of sociopathy being advantageous to the individual if population-wide frequencies of it are low, and vice versa. I argue that at least one alternative hypothesis exists that explains her data equally well. Alternative hypotheses must be formulated and tested before any theory can be validated.

In her target article, Mealey (1995) presented a comprehensive theory on the evolution of sociopathy. One of the pillars of her theory is the finding of significant heritabilities for sociopathy.[1] Because genetic variation for sociopathy is present in the population, her next step is then to hypothesize that it follows that sociopathy is subjected to frequency dependent selection. Sociopathy will be advantageous to the individual in question if the frequency of sociopathy is low in the population, and vice versa.

Mealey does not provide any alternative hypothesis,[2] and her whole theory is, in fact, an attempt to arrive at a post hoc explanation for a diverse number of observations. I intend to show here that alternative hypotheses, with vastly different implications, can sometimes be formulated easily. In short, hypothesis generation and testing urgently deserve more attention in sociobiological theorizing.

My argument is simple. Mealey hypothesizes a sort of temporal stabilizing selection for sociopathy, leading to a seesawing of its frequency in the population; depending on its frequency, sociopathy will confer reproductive advantages or disadvantages upon afflicted individuals. However, it would appear that a more classical form of stabilizing selection, constant over time, for intermediate levels of socialization would also explain the occurrence of sociopaths without the need to hypothesize that sociopathy is an advantageous evolutionary stable strategy (ESS) and could at any point be an advantageous reproductive strategy. Most or even all of Mealey's arguments are compatible with an interpretation where both extremely high and low levels of antisocial behavior would be disadvantageous, intermediate levels being most optimal. Such a type of stabilizing selection leads to a genetic architecture of large additive genetic effects and ambidirectional dominance[3] (Broadhurst & Jinks 1974).

It can easily be seen that such a genetic architecture would lead to a population composed mostly of individuals having intermediate levels for the phenotype upon which the stabilizing selection is acting. Alleles predisposing an individual for higher levels of ex-

https://doi.org/10.1017/S0140525X04210044 Published online by Cambridge University Press