When do representations explain?

Daniel C. Dennett

Philosophy Department, Tufts University, Medford, Mass. 02155

Stabler's patient and insightful clarification of this issue is most welcome, for he has provided some quite visible and stationary landmarks against which to detect the ideological slippage that has so far marked this debate. He is reticent, however, in answering one dialectically important question: if, as he convincingly argues, none of the evidence to date supports H3, what sort of evidence conceivably *could* support it?

For a representation to figure as a representation in a causal explanation, it must occur in a context where it is "read" by some agent, organ, or device. What could establish that such a process occurs? Consider an obvious case: old Mother Hubbard lies dead on the floor, a victim of poisoning, an open and half-full bottle of paint remover in the cupboard. Acquaintances say she had not been depressed, but had complained recently of "fainting spells." "Aha!" says the detective, noting her thick eyeglasses. "The bottle label says 'FOR PEELING PAINT' and she must have misread it to say 'FOR FEELING FAINT.' See how like Fs those Ps are."

What could conceivably convince us that actual rule consultation occurs in language processing would be evidence that on occasion rules are misread. But evidence for this would require some extraordinarily hard-to-acquire sorts of supporting evidence: evidence about not just the function or operation of the rules (as Stabler shows), and not even about just the "abstract" form of the rules (for, as Stabler shows, this evidence is always reinterpretable as evidence about function), but about the actual physical features of the encoding and the reading mechanisms – not just the semantics and syntax of the language of thought, but its orthography and typography as well. Could anything less give us clear evidence in support of H3 over its more modest rivals? So far as I can see, nothing else would be *direct* evidence.

The trouble is that it is not clear, given Stabler's treatment of the program/data distinction, that even this sort of evidence would satisfy him. For how could we distinguish, given this incredibly strong (imagined) evidence, the alternative hypothesis that we had not simply uncovered the typography of the data-representing system, rather than the program-reading system? I am inclined to conclude that there is something fishy about Stabler's attempt to make that distinction, at least as it would have to be adjusted to be transported from computerland to psycholinguistics. Consider another simple case: we teach somebody a simple algorithm for performing some congitive task, such as deciding whether to open the bidding in bridge, or winning at Nim. This, then, will be a paradigm case of someone – in this case consciously, even self-consciously – consulting a remembered rule and guiding calculation by its lights. Is it a case in which the rule counts as data – "the rules as argument" – or as program? Perhaps the answer is obvious, but it was not obvious to me what Stabler's answer would be.

Supposing this point clarified somehow, we might return to the question of whether there might be *indirect* evidence strongly supporting the existence of represented rules. One

possible line of argument, hinted at but skirted by Stabler, is one form or another of the "you can't get there from here" argument. One *might* argue, that is, that while "hybrid" and "hardwired" systems are always possible in principle and even, once created, faster and more efficient, they can only be created "naturally" by a design process that first implements a system in which the rules essential to the "rationale" of the system's function are explicit, and explicitly consulted. I think this is a risky and dubious sort of speculation, but its rationale is probably worth exploring. Consider the advanced bridge player who no longer consciously "counts points" (and who might not be counting them unconsciously either); there is surely some plausibility to the idea that the sophisticated but *ex hypothesi* merely H1 rule-described behavior of this player could only have been entrained by a process that includes an interim stage of H3 rule following. In a similar vein, one could argue that it is no accident that sophisticated hardwired microchips – such as those to be found in arcade video games – are designed by a process that begins with a program-guided system in which the operations are debugged. Tempting as these analogies are, however, they serve in the present context to highlight one of the most compelling sorts of indirect evidence *against* any H3-type theory of human linguistic competence. Surely the evolutionary design process that yields our innate linguistic competence as its product is strongly disanalogous to the design process that yields video games, precisely in being undirected, unforesighted, and completely lacking the sort of explicit "top-down" goal that is the hallmark of design (or training) processes that are aided by explicit "rules for beginners." [See also Dennett: "Intentional Systems in Cognitive Ethology" *BBS* 6(3) 1983.]