ACTIVITAS NERVOSA SUPERIOR

*Activitas Nervosa Superior 2009;**52**:2,81-84*

IDEAS & PERSPECTIVES

TWO BLACK BOXES: A FABLE[1]

Daniel Dennett*

Center for Cognitive Studies, Tufts University, Medford, MA, USA
Received June 23, 2010; accepted June 25, 2010

Abstract

Once upon a time, there were two large black boxes, A and B, connected by a long insulated copper wire. On box A there were two buttons, marked *a* and *b*, and on box B there were three lights, red, green, and amber. Scientists studying the behavior of the boxes had observed that whenever you pushed the *a* button on box A, the red light flashed briefly on box B, and whenever you pushed the *b* button on box A, the green light flashed briefly. The amber light never seemed to flash. They performed a few billion trials, under a very wide variety of conditions, and found no exceptions. There seemed to them to be a causal regularity.

Key words: Black box; Causality; Regularity

Once upon a time, there were two large black boxes, A and B, connected by a long insulated copper wire.[2] On box A there were two buttons, marked *a* and *b*, and on box B there were three lights, red, green, and amber. Scientists studying the behavior of the boxes had observed that whenever you pushed the *a* button on box A, the red light flashed briefly on box B, and whenever you pushed the *b* button on box A, the green light flashed briefly. The amber light never seemed to flash. They performed a few billion trials, under a very wide variety of conditions, and found no exceptions. There seemed to them to be a causal regularity, which they conveniently summarized thus:

all *a*'s cause reds

all *b*'s cause greens

The causation passed through the copper wire somehow, they determined, since severing it turned off all effects in box B, and shielding the two boxes from each other without severing the wire never disrupted the regularity. So naturally they were curious to know just how the causal regularity they had discovered was effected through the wire. Perhaps, they thought, pressing button *a* caused a low voltage pulse to be emitted down the wire, triggering the red light, and pressing button *b* caused a high voltage pulse, which triggered the green. Or perhaps pressing *a* caused a single pulse, which triggered the red light, and pressing *b* caused a double pulse. Clearly, there was something that always happened in the wire when you pressed button *a*, and something different that always happened in the wire when you pressed *b*. Discovering just what this was would explain the causal regularity they had discovered.

A wiretap of sorts on the wire soon revealed that things were more complicated. Whenever either button was pushed on box A, a long stream of pulses and gaps (ons and offs, or bits) was sent swiftly down the wire to box B--10,000 bits, to be exact. But it was a different pattern each time!

Clearly there had to be a feature or property of the strings of bits that triggered the red light in one case and the green light in the other. What could it be? They decided to open up box B and see what happened to the strings of bits when they arrived. Inside B they found a supercomputer--just an ordinary digital, serial supercomputer, with a large memory, containing a huge program and a huge data base, written, of course, in more bit strings. And when they traced the effects of the incoming bit strings on this computer program, they found nothing out of the ordinary: the input string would always make its way into the CPU in normal fashion, where it would provoke a few billion operations to be performed in a few seconds, ending, always, with either of two output signals, a 1 (which turned on the red light) or a 0 (which turned on the green light). In

**Correspondence to: Daniel Dennett, email: Daniel.Dennett@tufts.edu*

every case, they found, they could explain each step of the causation at the microscopic level without any difficulty or controversy. No occult causes were suspected to be operating, and, for instance, when they arranged to input the same sequence of 10,000 bits again and again, the program in box B always yielded the same output, red or green.

But this was mildly puzzling, because although it always gave the same output, it didn't always yield the same output by going through the same intermediate steps. In fact, it almost always passed through different states before yielding the same output. This in itself was no mystery, because the program kept a copy of each input it received, and so, when the same input arrived a second or third or thousandth time, the state of the memory of the computer was slightly different each time. But the output was always the same; if the light turned red the first time a particular string was input, it always turned red for the same string thereafter, and the same regularity held for green strings (as the scientists began to call them). All strings, they were tempted to hypothesize, are either red strings (cause the red light to flash) or green strings (cause the green light to flash). But of course they hadn't tested all possible strings--only strings that had been emitted by box A.

So they decided to test their hypothesis by disconnecting A from B temporarily and inserting variations on A's output strings to B. To their puzzlement and dismay, they discovered that almost always, when they tampered with a string from A, the amber light flashed! It was almost as if box B had detected their intervention. There was no doubt, however, that box B would readily accept man-made versions of red strings by flashing red, and man-made versions of green strings by flashing green. It was only when a bit--or more than one bit--was changed in a red or green string that the amber light usually--almost always--came on. A flurry of experimentation with a few billion random variations on bit-strings of length 10,000 strongly suggested to the scientists that there were really three varieties of strings: red strings, green strings, and amber strings--and amber strings outnumbered red and green strings by many, many orders of magnitude. Almost all strings were amber strings. That made the regularity they had discovered all the more exciting and puzzling.

What was it about red strings that turned on the red light and green strings that turned on the green light? Of course in each particular case, there was no mystery at all. They could trace the causation of each particular string through the supercomputer in B and see that, with gratifying Laplacean determinism, it produced its red or green or amber light, as the case might be. What they couldn't find, however, was a way of predicting which of the three effects a new string would have, just by examining it (without "hand-simulating" its effect on box B). They knew from their empirical data, of course, that the odds were very high that any new string considered would be amber--unless it was a string known to have been emitted by box A, in which case the odds were better than a billion to one that it would be either red or green, but no one could tell which, without running it through box B to see how the program settled out.

Since in spite of much brilliant and expensive research they found themselves still utterly unable to predict whether a string would turn out to be red, green, or amber, some theorists were tempted to call these properties emergent properties. What they meant was that the properties were (they thought) unpredictable in principle from a mere analysis of the microproperties of the strings themselves. But this didn't seem likely at all, since each particular case was as predictable as any deterministic input to a deterministic program could be. In any event, whether or not the properties of red, green and amber were unpredictable in principle, or merely in practice, they certainly were surprising and mysterious properties.

Perhaps the solution to the mystery lay in box A. They opened it up and found another supercomputer--of a different make and model, and running a different gigantic program, but also just a garden variety digital computer. They soon determined that whenever you pushed button *a* this sent the program off in one way, by sending a code (11111111) to the CPU, and whenever you pushed button *b* this sent a different code (00000000) to the CPU, setting in motion a different set of billions of operations. It turned out that there was an internal "clock" ticking away millions of times a second, and whenever you pushed either button the first thing the computer did was take the "time" from the clock (e.g., 10110101010101010111) and break it up into strings it then used to determine which subroutines to call in which order, and which part of its memory to access first in the course of its preparation of a bit string to send down the wire.

The scientists were able to figure out that it was this clock-consulting (which was as good as random) that virtually guaranteed that the same bit-string was never sent out twice. But in spite of this randomness, or pseudo-randomness, it remained true that whenever you pushed button *a*, the bit string the computer concocted turned out to be red, and whenever you pushed button *b*, the bit string eventually sent turned out to be green. Actually, the scientists did find a few anomalous cases: in roughly one in a billion trials, pushing the *a* button caused a green string to be emitted, or pushing the *b* button caused a red string to be emitted. This tiny blemish in perfection only whetted the scientists' appetite for an explanation of the regularity.

And then one day along came the two AI hackers who had built the boxes, and they explained it all. Al, who had built box A, had been working for years on an "ex-

pert system"--a data base containing "true propositions" about everything under the sun, and an inference engine to deduce further implications from the axioms that composed the data base. There were major league baseball statistics, meteorological records, biological taxonomies, histories of the world's nations, and hosts of trivia in the data base. Bo, the Swede who had build box B, had been working during the same time on a rival "world knowledge" data-base for his own expert system. They had each stuffed their respective data bases with as many "truths" as years of work had permitted[3].

But as the years progressed, they had grown bored with expert systems, and had both decided that the practical promise of this technology was vastly overrated. The systems weren't actually very good at solving interesting problems, or "thinking", or "finding creative solutions to problems." All they were good at, thanks to their inference engines, was generating lots and lots of true sentences (in their respective languages), and testing input sentences (in their respective languages) for truth and falsity--relative to their "knowledge" of course. So Al and Bo had gotten together and figured out how the fruits of their wasted effort could be put to use. They decided to make a philosophical toy. They chose a lingua franca for translating between their two representational systems (it was English, actually, sent in ASCII code), and hooked the machines together with a wire. Whenever you pushed A's *a* button, this instructed A to choose at random (or pseudo-random) one if its "beliefs" (either a stored axiom or a generated implication of its axioms), translate it into English, and send it to B, which translated this input into its own language (which was Swedish Lisp), and tested it against its own "beliefs"--its data base. Since both data bases were composed of truths, and roughly the same truths, thanks to their inference engines, whenever A sent B something A "believed," B "believed" it too, and signaled this by flashing a red light. Whenever A sent B what A took to be a falsehood, B announced that it judged that this was indeed a falsehood by flashing a green light. And whenever anyone tampered with the transmission, this almost always resulted in a string that was not a well-formed sentence of English (B had absolutely zero tolerance for "typographical" errors). B responded to these by flashing the amber light. Whenever anyone chose a bit string at random, the odds were high indeed that it would not be a well formed truth or falsehood of English ASCII, hence the preponderance of amber strings.

So, said Al and Bo, the emergent property red was actually the property of being a true sentence of English, and green was the property of being a falsehood in English. Suddenly, the search that had eluded the scientists for years became child's play. Anyone could compose red strings ad nauseam--just write down the ASCII code for "Houses are bigger than peanuts" or "Whales don't fly" or "Three times four is two less than two times seven" for instance. If you wanted a green string, try "Nine is less than eight" or "New York is the capital of Spain." Philosophers soon hit upon cute tricks, such as finding strings that were red the first hundred times they were given to B but green thereafter (e.g., the ASCII for "This sentence has been sent to you for evaluation less than a hundred and one times.")

But, said some philosophers, the string properties red and green are not really truth in English and falsity in English. After all, there are English truths whose ASCII expression takes millions of bits, and besides, Al and Bo didn't always insert facts in their programs. Some of what had passed for common knowledge when they were working on their data bases had since been disproven. And so forth. There were lots of reasons why the string property--the causal property--of redness was not quite exactly the property of truth in English. So, perhaps red could better be defined as relatively short expression in English ASCII of something 'believed' true by box B (whose 'beliefs' are almost all true). This satisfied some, but other picked nits, insisting, for various reasons, that this definition was inexact, or had counterexamples that could not be ruled out in any non-adhoc way, and so forth. But as Al and Bo pointed out, there were no better candidate descriptions of the property to be found, and hadn't the scientists been yearning for just such an explanation? Hadn't the mystery of red and green strings now been entirely dissolved? Moreover, now that it was dissolved, couldn't one see that there wasn't any hope at all of explaining the causal regularity with which we began our tale without using some semantical (or mentalistic) terms?

Some philosophers argued that while they had now indeed discovered a regularity in the activity in the wire that could be used to predict box B's behavior, it wasn't a causal regularity after all. Nonsense, others retorted. Pushing button *a* causes the red light to go on just as certainly as turning the ignition key causes your car to start. If it had turned out that what was being sent down the wire was simply high versus low voltage, or one pulse versus two, everybody would agree that this was a paradigm causal system. The fact that this system turned out to be a Rube Goldberg machine didn't show that the reliability of the link between *a* and red flashes was any less causal. In fact in every single case the scientists could trace out the exact micro-causal path that explained the result[4].

Convinced by this line of reasoning, other philosophers began to argue that this showed that the properties red,

green and amber weren't really semantical or mentalistic properties after all, but only imitation semantical properties, "as if" semantical properties. What red and green were, really, were very, very complicated syntactical properties. (They declined, however, to say anything further about just what syntactical properties these were, or to explain how even young children could swiftly and reliably produces instances of them, or recognize them.) "I suppose," retorted Al and Bo, "that if you had found us inside our black boxes, playing a trick on you by following the same scheme, you would then relent and agree that the operative causal property was genuine truth (or believed-truth, in any event). Can you propose any good reason for drawing such a distinction?" This led some to declare that in a certain important sense Al and Bo had been in the boxes, since they were responsible for creating the respective data bases, as models of their own beliefs. It led others to denying that there really were any semantical or mentalistic properties anywhere in the world. Content, they said, had been eliminated. The debate went on for years, but the mystery with which we began was solved.

(Pedantic Postscript)

I would like to end the tale there, and leave the rest for discussion, but various philosophers have insisted that I spell out what I take to be the implications of my fable. Here are a few:

1. Although there no doubt is some godforsaken and utterly inscrutable syntactic predicate that captures the feature of "red" and "green" strings (after all, the two boxes are paradigm cases of what I have called syntactic engines), there is really no substitute for semantic or intentional predicates when it comes to specifying the property in a compact, generative, explanatory way. Try it; you'll see I'm right.
2. The claim that if Al and Bo were discovered to be in their respective boxes, the situation would be importantly different, cannot be plausibly sustained. For surely there is some godforsaken and utterly inscrutable syntactic property that would also capture Bo's proclivity to turn on the red and green light in response to input strings; his brain is, after all, a syntactic engine. (Denying that lands you in a downright mystical doctrine.)
3. The situation described is not restricted to semantic predicates. Here are two other domains in which much the same morals could, and should, be drawn:

i. phoneme recognition (think of the task of specifying the physical properties shared by all acoustic signals detected as containing the English phoneme "r", for instance).

ii. color vision (what do all actually red and green things have in common?)

Endnotes

1. This fable originally appeared in Dennett, *Darwin's Dangerous Idea: Evolution and the Meanings of Life* (Dennett, 1995).
2. This began as an impromptu response to Jaegwon Kim's talk at Harvard, November 29, 1990: "Emergence, Non-Reductive Materialism, and 'Downward Causation'."
3. For a real-world example of such a project, see Douglas Lenat's enormous CYC (short for encyclopedia) project at MCC.
4. Some have argued that my account of patterns in "Real Patterns" (Dennett, 1991) is epiphenomenalism about content. This is my reply.

REFERENCES

Dennett, D. (1991). Real Patterns. *Journal of Philosophy 88*, 27-51.

Dennett, D. (1995). *Darwin's Dangerous Idea: Evolution and the Meanings of Life.* New York: Simon and Schuster.