THE MIND'S I

Fantasies and Reflections on Self and Soul

COMPOSED AND ARRANGED BY
Douglas R. Hofstadter
AND
Daniel C. Dennett

BANTAM BOOKS
TORONTO • NEW YORK • LONDON • SYDNEY • AUCKLAND
Minds, Brains, and Programs

not have any mental states. Whatever it is that the brain does to produce intentionality, it cannot consist in instantiating a program since no program, by itself, is sufficient for intentionality.*

Reflections

This article originally appeared together with twenty-eight responses from assorted people. Many of the responses contained excellent commentary, but reprinting them would have overloaded this book, and in any case some were a little too technical. One of the nice things about Searle’s article is that it is pretty much understandable by someone without special training in AI, neurology, philosophy, or other disciplines that have a bearing on it.

Our position is quite opposed to Searle’s, but we find in Searle an eloquent opponent. Rather than attempt to give a thorough rebuttal to his points, we will concentrate on a few of the issues he raises, leaving our answers to his other points implicit, in the rest of this book.

Searle’s paper is based on his ingenious “Chinese room thought experiment,” in which the reader is urged to identify with a human being executing by hand the sequence of steps that a very clever AI program would allegedly go through as it read stories in Chinese and answered questions about them in Chinese in a manner sufficiently human-seeming as to be able to pass the Turing test. We think Searle has committed a serious and fundamental misrepresentation by giving the impression that it makes any sense to think that a human being could do this. By buying this image, the reader is unwittingly sucked into an impossibly unrealistic concept of the relation between intelligence and symbol manipulation.

The illusion that Searle hopes to induce in readers (naturally he doesn’t think of it as an illusion!) depends on his managing to make readers overlook a tremendous difference in complexity between two systems at different conceptual levels. Once he has done that, the rest is a piece of cake. At the outset, the reader is invited to identify with Searle

* I am indebted to a rather large number of people for discussion of these matters and for their patient attempts to overcome my ignorance of artificial intelligence. I would especially like to thank Ned Block, Hubert Dreyfus, John Haugeland, Roger Schank, Robert Wilensky, and Terry Winograd.
as he hand-simulates an existing AI program that can, in a limited way, answer questions of a limited sort, in a few limited domains. Now, for a person to hand-simulate this, or any currently existing AI program—that is, to step through it at the level of detail that the computer does—would involve days, if not weeks or months, of arduous, horrendous boredom. But instead of pointing this out, Searle—as deit at distracting the reader's attention as a practiced magician—switches the reader’s image to a hypothetical program that passes the Turing test! He has jumped up many levels of competency without so much as a passing mention. The reader is again invited to put himself or herself in the shoes of the person carrying out the step-by-step simulation, and to “feel the lack of understanding” of Chinese. This is the crux of Searle’s argument.

Our response to this (and, as we shall show later, Searle’s response as well, in a way) is basically the “Systems Reply”: that it is a mistake to try to impute the understanding to the (incidentally) animate simulator; rather it belongs to the system as a whole, which includes what Searle casually characterizes as “bits of paper.” This offhand comment, we feel, reveals how Searle’s image has blinded him to the realities of the situation. A thinking computer is as repugnant to John Searle as non-Euclidean geometry was to its unwitting discoverer: Gerolamo Saccheri, who thoroughly disowned his own creation. The time—the late 1700s—was not quite ripe for people to accept the conceptual expansion caused by alternate geometries. About fifty years later, however, non-Euclidean geometry was rediscovered and slowly accepted.

Perhaps the same will happen with “artificial intentionality”—if it is ever created. If there ever came to be a program that could pass the Turing test, it seems that Searle, instead of marveling at the power and depth of that program, would just keep on insisting that it lacked some marvelous “causal powers of the brain” (whatever they are). To point out the vacuity of that notion, Zenon Pylshyn, in his reply to Searle, wondered if the following passage, quite reminiscent of Zuboff’s “Story of a Brain” (selection 12), would accurately characterize Searle’s viewpoint:

If more and more of the cells in your brain were to be replaced by integrated circuit chips, programmed in such a way as to keep the input-output function of each unit identical to that of the unit being replaced, you would in all likelihood just keep right on speaking exactly as you are doing now except that you would eventually stop meaning anything by it. What we outside observers might take to be words would become for you just certain noises that circuits caused you to make.

The weakness of Searle’s position is that he offers no clear way to tell when genuine meaning—or indeed the genuine “you”—has vanished from this system. He merely insists that some systems have intentionality by virtue of their “causal powers” and that some don’t. He vacillates about what those powers are due to. Sometimes it seems that the brain is composed of “the right stuff,” but other times it seems to be something else. It is whatever seems convenient at the moment—now it is the slippery essence that distinguishes “form” from “content,” now another essence that separates syntax from semantics, and so on.

To the Systems-Reply advocates, Searle offers the thought that the human being in the room (whom we shall from now on refer to as “Searle’s demon”) should simply memorize, or incorporate all the material on the “bits of paper.” As if a human being could, by any conceivable stretch of the imagination, do this. The program on those “bits of paper” embodies the entire mind and character of something as complex as its ability to respond to written material as a human being is, by virtue of being able to pass the Turing test. Could any human being simply “swallow up” the entire description of another human being’s mind? We find it hard enough to memorize a written paragraph; but Searle envisions the demon as having absorbed what in all likelihood would amount to millions, if not billions, of pages densely covered with abstract symbols—and moreover having all of this information available, whenever needed, with no retrieval problems. This unlikely aspect of the scenario is all lightly described, and it is not part of Searle’s key argument to convince the reader that it makes sense. In fact, quite the contrary—a key part of his argument is in glossing over these questions of orders of magnitude, for otherwise a skeptical reader will realize that nearly all of the understanding must lie in the billions of symbols on paper, and practically none of it in the demon. The fact that the demon is animate is an irrelevant—indeed, misleading—side issue that Searle has mistaken for a very significant fact.

We can back up this argument by exhibiting Searle’s own espousal of the Systems Reply. To do so, we should first like to place Searle’s thought experiment in a broader context. In particular, we would like to show how Searle’s setup is just one of a large family of related thought experiments, several of which are the topics of other selections in this book. Each member of this family of thought experiments is defined by a particular choice of “knob settings” on a thought-experiment generator. Its purpose is to create—in your mind’s eye—various sorts of imaginary simulations of human mental activity. Each different thought experiment is an “intuition pump” (Dennett’s term) that magnifies one facet or other of the issue, tending to push the reader toward certain conclusions. We see approximately five knobs of interest, although it is possible that someone else could come up with more.
Knob 1. This knob controls the physical “stuff” out of which the simulation will be constructed. Its settings include: neurons and chemicals; water pipes and water; bits of paper and symbols on them; toilet paper and stones; data structures and procedures; and so on.

Knob 2. This knob controls the level of accuracy with which the simulation attempts to mimic the human brain. It can be set at an arbitrarily fine level of detail (particles inside atoms), at a coarser level such as that of cells and synapses, or even at the level that AI researchers and cognitive psychologists deal with: that of concepts and ideas, representations and processes.

Knob 3. This knob controls the physical size of the simulation. Our assumption is that microminiaturization would allow us to make a teeny-weeny network of water pipes or solid-state chips that would fit inside a thimble, and conversely that any chemical process could be blown up to the macroscopic scale.

Knob 4. This critical knob controls the size and nature of the demon who carries out the simulation. If it is a normal-sized human being, we shall call it a “Searle’s demon.” If it is a tiny elflike creature that can sit inside neurons or on particles, then we shall call it a “Haugeland’s demon,” after John Haugeland, whose response to Searle featured this notion. The settings of this knob also determine whether the demon is animate or inanimate.

Knob 5. This knob controls the speed at which the demon works. It can be set to make the demon work blindingly fast (millions of operations per microsecond) or agonizingly slowly (maybe one operation every few seconds).

Now, by playing with various knob settings, we can come up with various thought experiments. One choice yields the situation described in selection 26, “A Conversation with Einstein’s Brain.” Another choice yields Searle’s Chinese room experiment. In particular, that involves the following knob settings:

Knob 1: paper and symbols
Knob 2: concepts and ideas
Knob 3: room size
Knob 4: human-sized demon
Knob 5: slow setting (one operation every few seconds)

Note that in principle Searle is not opposed to assuming that a simulation with these parameters could pass the Turing test. His dispute is only with what that would imply.

There is one final parameter that is not a knob but a point of view from which to look at the experiment. Let us add a little color to this drab experiment and say that the simulated Chinese speaker involved is a woman and that the demons (if animate) are always male. Now we have a choice between the demon’s-eye view and the system’s-eye view. Remember that by hypothesis, both the demon and the simulated woman are equally capable of articulating their views on whether or not they are understanding, and on what they are experiencing. Searle is insistent, nonetheless, that we see this experiment only from the point of view of the demon. He insists that no matter what the simulated woman claims (in Chinese, of course) about her understanding, we should disregard her claims, and pay attention to the demon inside, who is carrying out the symbol manipulation. Searle’s claim amounts to the notion that actually there is only one point of view, not two. If one accepts the way Searle describes the whole experiment, this claim has great intuitive appeal, since the demon is about our size, speaks our language, and works at about our speed—and it is very hard to identify with a “woman” whose answers come at the rate of one per century (with luck)—and in “meaningless squiggles and squiggles,” to boot.

But if we change some of the knob settings, we can also alter the ease with which we change point of view. In particular, Haugeland’s variation involves switching various knobs as follows:

Knob 1: neurons and chemicals
Knob 2: neural-firing level
Knob 3: brain size
Knob 4: eensy-weensy demon
Knob 5: dazzlingly fast demon

What Haugeland wants us to envision is this: A real woman’s brain is, unfortunately, defective. It no longer is able to send neurotransmitters from one neuron to another. Luckily, however, this brain is inhabited by an incredibly tiny and incredibly speedy Haugeland’s demon, who intervenes every single time any neuron would have been about to release neurotransmitters into a neighboring neuron. This demon “tickles” the appropriate synapse of the next neuron in a way that is functionally indistinguishable, to that neuron, from the arrival of genuine neurotransmitters. And the H-demon is so swift that he can jump around from synapse to synapse in trillionths of a second, never falling behind schedule. In this way the operation of the woman’s brain proceeds exactly as it would have, if she were healthy. Now, Haugeland asks Searle, does the woman still think—that is, does she possess intentionality—or, to recall the words of Professor Jefferson as cited by Turing, does she merely “artificially signal”?
You might expect Searle to urge us to listen to and identify with the
demon, and to eschew the Systems Reply, which would be, of course, to
listen to and identify with the woman. But in his response to Haugeland,
Searle surprises us—he chooses to listen to her this time and to ignore
the demon who is cursing us from his tiny vantage point, yelling up to
us, “Fools! Don’t listen to her! She’s merely a puppet whose every action
is caused by my tickling, and by the program embedded in these many
neurons that I zip around among.” But Searle does not heed the H-
demon’s warning cries. He says, “Her neurons still have the right causal
powers; they just need some help from the demon.”

We can construct a mapping between Searle’s original setup and this
modified setup. To the “bits of paper” now correspond all the
synapses in the woman’s brain. To the AI program written on these
“bits of paper” corresponds the entire configuration of the woman’s
brain; this amounts to a gigantic prescription telling the demon when and
how to know which synapses to tickle. To the act of writing “meaningless
squiggles and squiggles of Chinese” on paper corresponds the act of
tickling her synapses. Suppose we take the setup as is, except that we’ll
vary the size and speed knobs. We’ll blow the woman’s brain up to the
size of the Earth, so that the demon becomes an “us-sized” S-demon
instead of a tiny H-demon. And let’s also have the S-demon act at speeds
reasonable for humans, instead of zipping thousands of miles throughout
this bulbous brain in mere microseconds. Now which level does Searle
wish us to identify with? We won’t speculate, but it seems to us that if the
Systems Reply was compelling in the previous case, it should still be so in
this case.

It must be admitted that Searle’s thought experiment vividly raises
the question of what understanding a language really is. We would like
to digress for a moment on that topic. Consider the question: “What kind
of ability to manipulate the written or spoken symbols of a language
amounts to a true understanding of that language?” Parrots who parrot
English do not understand English. The recorded voice of a woman
announcing the exact time of day on the telephone time service is not the
mouthpiece of a system that understands English. There is no mentality
behind that voice—it has been skimmed off of its mental substrate, yet
retains a human-seeming quality. Perhaps a child would wonder how
anyone could have so boring a job, and could do it so reliably. This would
amuse us. It would be another matter, of course, if her voice were being
driven by a flexible AI program that could pass the Turing test!

Imagine you are teaching a class in China. Further, imagine that you
are aware of formulating all your thoughts in English and then of applying
last-minute transformation rules (in reality, they would be last-split-
second rules) that convert the English thoughts into instructions for
moving your mouth and vocal cords in strange, “meaningless” ways—
and yet, all your pupils sit there and seem quite satisfied with your
performance. When they raise their hands, they utter exotic sounds that,
although they are completely meaningless to you, you are equipped to
deal with, as you quickly apply some inverse rules and recover the English
meanings underlying them.... Would you feel you were actually speaking
Chinese? Would you feel you had gained some insight into the Chinese
mentality? Or—can you actually imagine this situation? Is it realistic?
Could anyone actually speak a foreign language well using this
method?

The standard line is “You must learn to think in Chinese.” But in what
does this consist? Anyone who has experienced it will recognize this
description: The sounds of the second language pretty soon become
“unheard”—you hear right through them, rather than hearing them, as
you see right through a window, rather than seeing the window. Of
course, you can make yourself hear a familiar language as pure uninter-
preted sound if you try very hard, just as you can look at a windowpane
if you want; but you can’t have your cake and eat it too—you can’t hear
the sounds both with and without their meanings. And so most of the time
people hear mainly meaning. For those people who learn a language
because of enchantment with its sounds, this is a bit disappointing—and
yet mastery of those sounds, even if one no longer hears them naively,
is a beautiful, exhilarating experience. (It would be an interesting thing
to try to apply this same kind of analysis to the hearing of music, where
the distinction between hearing bare sounds and hearing their “mean-
ings” is far less well understood, yet seems very real.)

Learning a second language involves transcending one’s own native
language. It involves mixing the new language right in with the medium
in which thought takes place. Thoughts must be able to germinate as
easily (or nearly as easily) in the new language as in one’s native language.
The way in which a new language’s habits seep down level by level and
finally get absorbed into neurons is a giant mystery still. But one thing
for certain is that mastery of a language does not consist in getting your
“English subsystem” to execute for you a program of rules that enable
you to deal with a language as a set of meaningless sounds and marks.
 Somehow, the new language must fuse with your internal represent-
tional system—your repertoire of concepts, images, and so on—in the
same intimate way as English is fused with it. To think precisely about
this, one must develop a very clear notion of the concept of levels of
implementation, a computer-science concept of great power.

Computer scientists are used to the idea that one system can “emu-
late" another system. In fact, it follows from a theorem proven in 1936 by Alan Turing that any general-purpose digital computer can take on the guise of any other general-purpose digital computer, and the only difference to the outside world will be one of speed. The verb "emulate" is reserved for simulations, by a computer, of another computer, while "simulate" refers to the modeling of other phenomena, such as hurricanes, population curves, national elections, or even computer users.

A major difference is that simulation is almost always approximate, depending on the nature of the model of the phenomenon in question, whereas emulation is in a deep sense exact. So exact is it that when, say, a Sigma-5 computer emulates a computer with different architecture—say a DEC PDP-10—the users of the machine will be unaware that they are not dealing with a genuine DEC. This embedding of one architecture in another gives rise to so-called "virtual machines"—in this case, a virtual DEC-10. Underneath every virtual machine there is always some other machine. It may be a machine of the same type, it may even be another virtual machine. In his book *Structured Computer Organization*, Andrew Tanenbaum uses this notion of virtual machines to explain how large computer systems can be seen as a stack of virtual machines implemented one on top of the other—the bottommost one being, of course, a real machine! But in any case, the levels are sealed off from each other in a watertight way, just as Searle's demon was prevented from talking to the Chinese speaker he was part of. (It is intriguing to imagine what kind of conversation would take place—assuming that there were an interpreter present, since Searle's demon knows no Chinese!)

Now in theory, it is possible to have any two such levels communicate with each other, but this has traditionally been considered bad style; level-mingling is forbidden. Nonetheless, it is probable that this forbidden fruit—this blurring of two implementational levels—is exactly what goes on when a human "system" learns a second language. The second language does not run on top of the first one as a kind of software parasite, but rather becomes equally fundamentally implanted in the hardware (or nearly so). Somehow, absorption of a second language involves bringing about deep changes in one's underlying "machine"—a vast and coherent set of changes in the ways that neurons fire, so sweeping a set of changes that it creates new ways for the higher-level entities—the symbols—to trigger one another.

To parallel this in a computer system, a higher-level program would have to some way of creating changes inside the "demon" that is carrying its program out. This is utterly foreign to the present style in computer science of implementing one level above another in a strictly vertical, sealed-off fashion. The ability of a higher level to loop back and affect lower levels—its own underpinnings—is a kind of magic trick which we feel is very close to the core of consciousness. It will perhaps one day prove to be a key element in the push toward ever-greater flexibility in computer design, and of course in the approach toward artificial intelligence. In particular, a satisfactory answer to the question of what "understanding" really means will undoubtedly require a much sharper delineation of the ways in which different levels in a symbol-manipulating system can depend on and affect one another. All in all, these concepts have proven elusive, and a clear understanding of them is probably a good way off yet.

In this rather confusing discussion of many levels, you may have started to wonder what in the world "level" really means. It is a most difficult question. As long as levels are sealed off from each other, like Searle's demon and the Chinese-speaking woman, it is fairly clear. When they begin to blur, beware! Searle may admit that there are two *levels* in his thought experiment, but he is reluctant to admit that there are two occupied *points of view*—two genuine beings that feel and "have experience." He is worried that once we admit that some computational systems might have experiences, that would be a Pandora's box and all of a sudden "mind would be everywhere"—in the churning of stomachs, livers, automobile engines, and so on.

Searle seems to believe that any system whatsoever can be ascribed beliefs and feelings and so on, if one looks hard enough for a way to describe the system as an instantiation of an AI program. Obviously, that would be a disturbing notion, leading the way to panpsychism. Indeed, Searle believes that the AI people have unwittingly committed themselves to a panpsychic vision of the world.

Searle's escape from his self-made trap is to maintain that all those "beliefs" and "feelings" that you will uncover in inanimate objects and so forth when you begin seeing mind everywhere are not genuine but "pseudo." They lack intentionality! They lack the causal powers of the brain! (Of course, Searle would caution others to beware of confusing these notions with the naively dualistic notion of "soul").

Our escape is to deny that the trap exists at all. It is incorrect to see minds everywhere. We say: minds do not lurk in car engines or livers any more than brains lurk in car engines and livers.

It is worthwhile expanding on this a little. If you can see all the complexity of thought processes in a churning stomach, then what's to prevent you from reading the pattern of bubbles in a carbonated beverage as coding for the Chopin piano concerto in E minor? And don't the holes in pieces of Swiss cheese code for the entire history of the United States? Sure they do—in Chinese as well as in English. After all, all things
are written everywhere! Bach's Brandenburg concerto no. 2 is coded for
in the structure of Hamlet—and Hamlet was of course readable (if you’d
only known the code) from the structure of the last piece of birthday cake
you gobbled down.

The problem is, in all these cases, that of specifying the code without
knowing in advance what you want to read. For otherwise, you could pull
a description of anyone’s mental activity out of a baseball game or a blade
of grass by an arbitrarily constructed a posteriori code. But this is not
science.

Minds come in different grades of sophistication, surely, but minds
worth calling minds exist only where sophisticated representational sys-
tems exist, and no describable mapping that remains constant in time will
reveal a self-updating representational system in a car engine or a liver.
Perhaps one could read mentality into a rumbling car engine in somewhat
the way that people read extra meanings into the structures of the Great
Pyramids or Stonehenge, the music of Bach, Shakespeare’s plays, and so
on—namely, by fabricating far-fetched numerological mapping schemes
that can be molded and flexed whenever needed to fit the desires of the
interpreter. But we doubt that that is what Scarle intends (we do grant
that he intends).

Minds exist in brains and may come to exist in programmed ma-
chines. If and when such machines come about, their causal powers will
derive not from the substances they are made of, but from their design
and the programs that run in them. And the way we will know they have
those causal powers is by talking to them and listening carefully to what
they have to say.

D.R.H.

Once upon a time there was a dualist. He believed that mind and matter
are separate substances. Just how they interacted he did not pretend to
know—this was one of the “mysteries” of life. But he was sure they were
quite separate substances.

This dualist, unfortunately, led an unbearably painful life—not be-
cause of his philosophical beliefs, but for quite different reasons. And he
had excellent empirical evidence that no respite was in sight for the rest
of his life. He longed for nothing more than to die. But he was deterred
from suicide by such reasons as: (1) he did not want to hurt other people
by his death; (2) he was afraid suicide might be morally wrong; (3) he was
afraid there might be an afterlife, and he did not want to risk the possibility
of eternal punishment. So our poor dualist was quite desperate.

Then came the discovery of the miracle drug! Its effect on the taker
was to annihilate the soul or mind entirely but to leave the body function-
ing exactly as before. Absolutely no observable change came over the
taker; the body continued to act just as if it still had a soul. Not the closest
friend or observer could possibly know that the taker had taken the drug,
unless the taker informed him.

Do you believe that such a drug is impossible in principle? Assuming
you believe it possible, would you take it? Would you regard it as im-
oral? Is it tantamount to suicide? Is there anything in Scriptures forbid-

"An Unfortunate Dualist" from This Book Needs No Title by Raymond M. Smullyan. Copy-
right © 1980 by Raymond M. Smullyan. Published by Prentice-Hall, Inc., Englewood Cliffs,
N.J.