16 When HAL Kills, Who's to Blame? Computer Ethics

Daniel C. Dennett

The first robot homicide was committed in 1981, according to my files. I have a yellowed clipping dated December 9, 1981, from the *Philadelphia Inquirer*—not the *National Enquirer*—with the headline "Robot killed repairman, Japan reports".

The story was an anticlimax. At the Kawasaki Heavy Industries plant in Akashi, a malfunctioning robotic arm pushed a repairman against a gearwheel-milling machine, which crushed him to death. The repairman had failed to follow instructions for shutting down the arm before he entered the workspace. Why, indeed, was this industrial accident in Japan reported in a Philadelphia newspaper? Every day somewhere in the world a human worker is killed by one machine or another. The difference, of course, was that—in the public imagination at least—this was no ordinary machine. This was a robot, a machine that might have a mind, might have evil intentions, might be capable, not just of homicide, but of murder. Anglo-American jurisprudence speaks of mens rea—literally, the guilty mind:

> To have performed a legally prohibited action, such as killing another human being; one must have done so with a culpable state of mind, or mens rea. Such culpable mental states are of three kinds: they are either motivational states of purpose, cognitive states of belief, or the nonmental state of negligence. (*Cambridge Dictionary of Philosophy,* 1995, p. 482)

The legal concept has no requirement that the agent be capable of feeling guilt or remorse or any other emotion; so-called cold-blooded murderers are

not in the slightest degree exculpated by their flat affective state. *Star Trek*'s Spock would fully satisfy the mens rea requirement in spite of his fabled lack of emotions. Drab, colorless—but oh so effective—"motivational states of purpose" and "cognitive states of belief" are enough to get the fictional Spock through the day quite handily. And they are well-established features of many existing computer programs.

When IBM's computer Deep Blue beat world chess champion Garry Kasparov in the first game of their 1996 championship match, it did so by discovering and executing, with exquisite timing, a withering attack, the purposes of which were all too evident in retrospect to Kasparov and his handlers. It was Deep Blue's sensitivity to those purposes and a cognitive capacity to recognize and exploit a subtle flaw in Kasparov's game that explain Deep Blue's success. Murray Campbell, Feng-hsiung Hsu, and the other designers of Deep Blue, didn't beat Kasparov; Deep Blue did. Neither Campbell nor Hsu discovered the winning sequence of moves; Deep Blue did. At one point, while Kasparov was mounting a ferocious attack on Deep Blue's king, nobody but Deep Blue figured out that it had the time and security it needed to knock off a pesky pawn of Kasparov's that was out of the action but almost invisibly vulnerable. Campbell, like the human grandmasters watching the game, would never have dared consider such a calm mopping-up operation under pressure.

Deep Blue, like many other computers equipped with artificial intelligence (AI) programs, is what I call an intentional system: its behavior is predictable and explainable if we attribute to it beliefs and desires—"cognitive states" and "motivational states"—and the rationality required to figure out what it ought to do in the light of those beliefs and desires. Are these skeletal versions of human beliefs and desires sufficient to meet the mens rea requirement of legal culpability? Not quite, but, if we restrict our gaze to the limited world of the chess board, it is hard to see what is missing. Since cheating is literally unthinkable to a computer like Deep Blue, and since there are really no other culpable actions available to an agent restricted to playing chess, nothing it could do would be a misdeed deserving of blame, let alone a crime of which we might convict it. But we also assign responsibility to agents in

order to praise or honor the appropriate agent. Who or what, then, deserves the credit for beating Kasparov? Deep Blue is clearly the best candidate. Yes, we may join in congratulating Campbell, Hsu and the IBM team on the success of their handiwork; but in the same spirit we might congratulate Kasparov's teachers, handlers, and even his parents. And, no matter how assiduously they may have trained him, drumming into his head the importance of one strategic principle or another, they didn't beat Deep Blue in the series: Kasparov did.

Deep Blue is the best candidate for the role of responsible opponent of Kasparov, but this is not good enough, surely, for full moral responsibility. If we expanded Deep Blue's horizons somewhat, it could move out into the arenas of injury and benefit that we human beings operate in. It's not hard to imagine a touching scenario in which a grandmaster deliberately (but oh so subtly) throws a game to an opponent, in order to save a life, avoid humiliating a loved one, keep a promise, or . . . (make up your own O'Henry story here). Failure to rise to such an occasion might well be grounds for blaming a human chess player. Winning or throwing a chess match might even amount to commission of a heinous crime (make up your own Agatha Christie story here). Could Deep Blue's horizons be so widened?

Deep Blue is an intentional system, with beliefs and desires about its activities and predicaments on the chessboard; but in order to expand its horizons to the wider world of which chess is a relatively trivial part, it would have to be given vastly richer sources of "perceptual" input—and the means of coping with this barrage in real time. Time pressure is, of course, already a familiar feature of Deep Blue's world. As it hustles through the multidimensional search tree of chess, it has to keep one eye on the clock. Nonetheless, the problems of optimizing its use of time would increase by several orders of magnitude if it had to juggle all these new concurrent projects (of simple perception and self-maintenance in the world, to say nothing of more devious schemes and opportunities). For this hugely expanded task of resource management, it would need extra layers of control above and below its chess-playing software. Below, just to keep its perceptuo-locomotor projects in basic coordination, it would need to have a set of rigid traffic-control

policies embedded in its underlying operating system. Above, it would have to be able to pay more attention to features of its own expanded resources, being always on the lookout for inefficient habits of thought, one of Douglas Hofstadter's "strange loops," obsessive ruts, oversights, and deadends. In other words, it would have to become a higher-order intentional system, capable of framing beliefs about its own beliefs, desires about its desires, beliefs about its fears about its thoughts about its hopes, and so on.

Higher-order intentionality is a necessary precondition for moral responsibility, and Deep Blue exhibits little sign of possessing such a capability. There is, of course, some self-monitoring implicated in any well-controlled search: Deep Blue doesn't make the mistake of reexploring branches it has already explored, for instance; but this is an innate policy designed into the underlying computational architecture, not something under flexible control. Deep Blue can't converse with you—or with itself—about the themes discernible in its own play; it's not equipped to notice—and analyze, criticize, analyze, and manipulate—the fundamental parameters that determine its policies of heuristic search or evaluation. Adding the layers of software that would permit Deep Blue to become self-monitoring and self-critical, and hence teachable, in all these ways would dwarf the already huge Deep Blue programming project—and turn Deep Blue into a radically different sort of agent.

HAL purports to be just such a higher-order intentional system—and he even plays a game of chess with Frank. HAL is, in essence, an enhancement of Deep Blue equipped with eyes and ears and a large array of sensors and effectors distributed around *Discovery 1*. HAL is not at all garrulous or self-absorbed; but in a few speeches he does express an interesting variety of higher-order intentional states, from the most simple to the most devious.

HAL: Yes, it's puzzling. I don't think I've ever seen anything quite like this before.

HAL doesn't just respond to novelty with a novel reaction; he notices that he is encountering novelty, a feat that requires his memory to have

an organization far beyond that required for simple conditioning to novel stimuli.

HAL: I can't rid myself of the suspicion that there are some extremely odd things about this mission.

HAL: I never gave these stories much credence, but particularly in view of some of the other things that have happened, I find them difficult to put out of my mind.

HAL has problems of resource management not unlike our own. Obtrusive thoughts can get in the way of other activities. The price we pay for adding layers of flexible monitoring, to keep better track of our own mental activities, is . . . more mental activities to keep track of!

HAL: I've still got the greatest enthusiasm and confidence in the mission. I want to help you.

Another price we pay for higher-order intentionality is the opportunity for duplicity, which comes in two flavors: self-deception and other-deception. Friedrich Nietzsche recognizes this layering of the mind as the key ingredient of the moral animal; in his overheated prose it becomes the "priestly" form of life:

> For with the priests everything becomes more dangerous, not only cures and remedies, but also arrogance, revenge, acuteness, profligacy, love, lust to rule, virtue, disease—but it is only fair to add that it was on the soil of this essentially dangerous form of human existence, the priestly form, that man first became an interesting animal, that only here did the human soul in a higher sense acquire depth and become evil—and these are the two basic respects in which man has hitherto been superior to other beasts! (*The Genealogy of Morals,* First Essay, 6)

HAL's declaration of enthusiasm is nicely poised somewhere between sincerity and cheap, desperate, canned ploy—just like some of the most important declarations we make to each other. Does HAL mean it? Could he mean it? The cost of being the sort of being that could mean it is the chance that he might not mean it. HAL is indeed an "interesting animal."

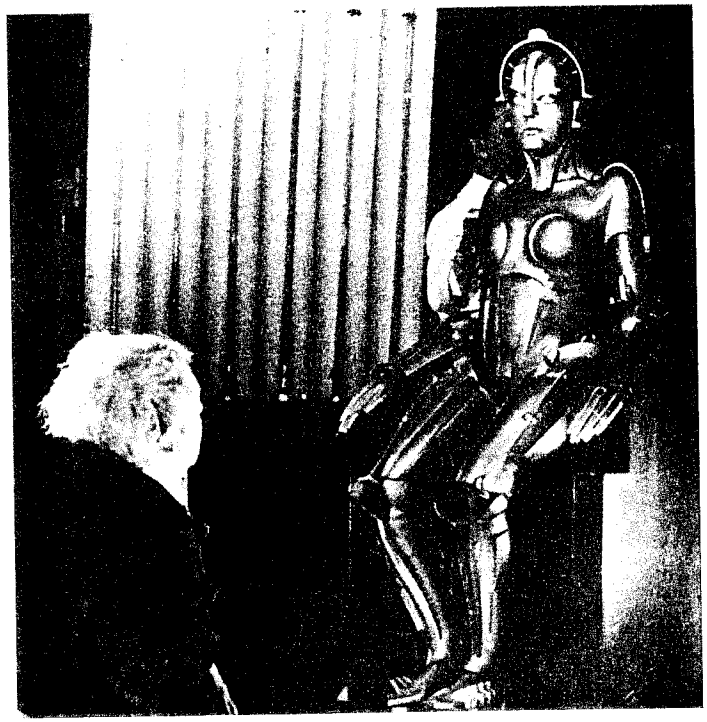

Figure 16.1
A Scene from Fritz Lang's Film *Metropolis* (1926)
Lang's robot is the beautiful but diabolical Maria.

But is HAL even remotely possible? In the book *2001,* Clarke has Dave reflect on the fact that HAL, whom he is disconnecting, "is the only conscious creature in my universe." From the omniscient-author perspective, Clarke writes about what it is like to be HAL.

> He was only aware of the conflict that was slowly destroying his integrity—the conflict between truth, and concealment of truth. He had begun to make mistakes, although, like a neurotic who could not observe his own symptoms, he would have denied it (p. 148).

Is Clarke helping himself here to more than we should allow him? Could something like HAL—a conscious, computer-bodied intelligent agent—be

brought into existence by any history of design, construction, training, learning, and activity? The different possibilities have been explored in familiar fiction and can be nested neatly in order of their descending "humanness."

1. *The Wizard of Oz.* HAL isn't a computer at all. He is actually an ordinary flesh-and-blood man hiding behind a techno-facade—the ultimate homunculus, pushing buttons with ordinary fingers, pulling levers with ordinary hands, looking at internal screens and listening to internal alarm buzzers. (A variation on this theme is John Searle's busy-fingered hand-simulation of the Chinese Room by following billions of instructions written on slips of paper.)

2. *William* (from "William and Mary," in *Kiss, Kiss* by Roald Dahl). HAL is a human brain kept alive in a "vat" by a life-support system and detached from its former body, in which it acquired a lifetime of human memory, hankerings, attitudes, and so forth. It is now harnessed to huge banks of prosthetic sense organs and effectors. (A variation on this theme is poor Yorick, the brain in a vat in the story, "Where Am I?" in my *Brainstorms.*)

3. *Robocop,* disembodied and living in a "vat." Robocop is part-human brain, part computer. After a gruesome accident, the brain part (vehicle of some of the memory and personal identity, one gathers, of the flesh-and-blood cop who was Robocop's youth) was reembodied with robotic arms and legs, but also (apparently) partly replaced or enhanced with special-purpose software and computer hardware. We can imagine that HAL spent some transitional time as Robocop before becoming a limbless agent.

4. *Max Headroom,* a virtual machine, a software duplicate of a real person's brain (or mind) that has somehow been created by a brilliant hacker. It has the memories and personality traits acquired in a normally embodied human lifetime but has been off-loaded from all-carbon-based hardware into a silicon-chip implementation. (A variation on this theme is poor Hubert, the software duplicate of Yorick, in "Where Am I?")

5. *The real-life but still-in-the-future*—and hence still strictly science-fictional—Cog, the humanoid robot being constructed by Rodney Brooks,

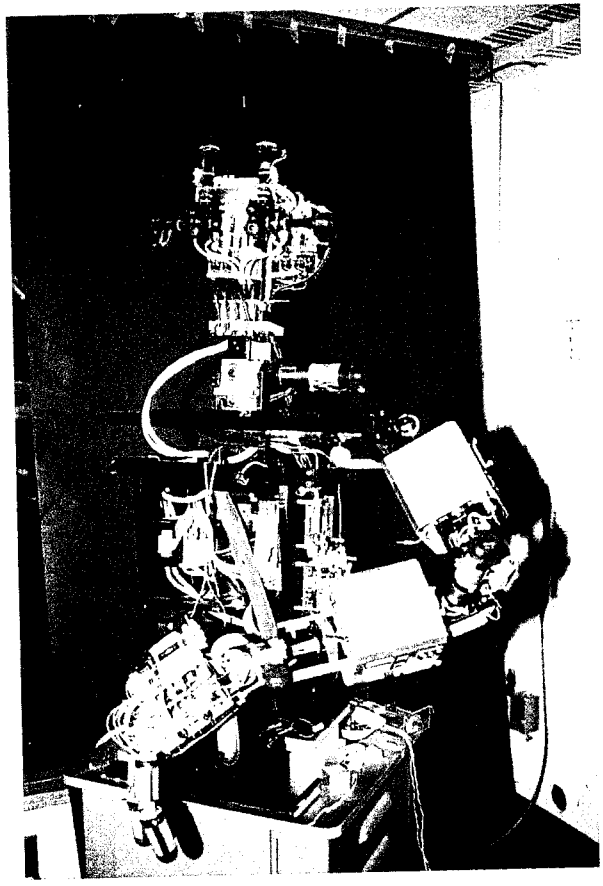

Figure 16.2
Cog, a Humanoid Robot Being Constructed at the MIT Artificial Intelligence Lab
The project is headed by Rodney Brooks, Lynn Andrea Stein, and Daniel C. Dennett. (Photo courtesy of the MIT Artificial Intelligence Lab)

Lynn Stein, and the Cog team at MIT (see figure 16.2). Cog's brain is all silicon chips from the outset, and its body parts are inorganic artifacts. Yet it is designed to go through an embodied infancy and childhood, reacting to people that it sees with its video eyes, making friends, learning about the world by playing with real things with its real hands, and acquiring memory. If Cog ever grows up, it could surely abandon its body and make the transi-

tion described in the fictional cases. It would be easier for Cog, who has always been a silicon-based, digitally encoded intelligence, to move into a silicon-based vat than it would be for Max Headroom or Robocop, who spent their early years in wetware. Many important details of Cog's degree of humanoidness (humanoidity?) have not yet been settled, but the scope is wide. For instance, the team now plans to give Cog a virtual neuroendocrine system, with virtual hormones spreading and dissipating through its logical spaces.

6. *Blade Runner in a vat* has never had a real humanoid body, but has hallucinatory memories of having had one. This entirely bogus past life has been constructed by some preposterously complex and detailed programming.

7. *Clarke's own scenario,* as best it can be extrapolated from the book and the movie. HAL has never had a body and has no illusions about his past. What he knows of human life he knows as either part of his innate heritage (coded, one gathers, by the labors of many programmers, after the fashion of the real-world CYC project of Douglas Lenat [see chapter 9]) or a result of his subsequent training—a sort of bedridden infancy, one gathers, in which he was both observer and, eventually, participant. (In the book, Clarke speaks of "the perfect idiomatic English he had learned during the fleeting weeks of his electronic childhood.")

The extreme cases at both poles are impossible, for relatively boring reasons. At one end, neither the Wizard of Oz nor John Searle could do the necessary handwork fast enough to sustain HAL's quick-witted round of activities. At the other end, hand-coding enough world knowledge into a disembodied agent to create HAL's dazzlingly humanoid competence and getting it to the point where it could benefit from an electronic childhood is a programming task to be measured in hundreds of efficiently organized person-centuries. In other words, the daunting difficulties observable at both ends of this spectrum highlight the fact that there is a colossal design job to be done; the only practical way of doing it is one version or another of Mother Nature's way—years of embodied learning. The trade-offs between various combinations of flesh-and-blood and silicon-and-metal bodies are

anybody's guess. I'm putting my bet on Cog as the most likely developmental platform for a future HAL.

Notice that requiring HAL to have a humanoid body and live concretely in the human world for a time is a practical but not a metaphysical requirement. Once all the R & D is accomplished in the prototype, by the odyssey of a single embodied agent, the standard duplicating techniques of the computer industry could clone HALs by the thousands as readily as they do compact discs. The finished product could thus be captured in some number of terabytes of information. So, in principle, the information that fixes the design of all those chips and hard-wired connections and configures all the RAM and ROM could be created by hand. There is no finite bit-string, however long, that is officially off-limits to human authorship. Theoretically, then, Blade-Runner-like entities could be created with ersatz biographies; they would have exactly the capabilities, dispositions, strengths, and weaknesses of a real, not virtual, person. So whatever moral standing the latter deserved should belong to the former as well.

The main point of giving HAL a humanoid past is to give him the world knowledge required to be a moral agent—a necessary modicum of understanding or empathy about the human condition. A modicum will do nicely; we don't want to hold out for too much commonality of experience. After all, among the people we know, many have moral responsibility in spite of their obtuse inability to imagine themselves into the predicaments of others. We certainly don't exculpate male chauvinist pigs who can't see women as people!

When *do* we exculpate people? We should look carefully at the answers to this question, because HAL shows signs of fitting into one or another of the exculpatory categories, even though he is a conscious agent. First, we exculpate people who are insane. Might HAL have gone insane? The question of his capacity for emotion—and hence his vulnerability to emotional disorder—is tantalizingly raised by Dave's answer to Mr. Amer.

Dave: Well, he acts like he has genuine emotions. Of course, he's pro-

grammed that way, to make it easier for us to talk to him. But as to whether he has real feelings is something I don't think anyone can truthfully answer.

Certainly HAL proclaims his emotional state at the end: "I'm afraid. I'm afraid." Yes, HAL is "programmed that way"—but what does that mean? It could mean that HAL's verbal capacity is enhanced with lots of canned expressions of emotional response that get grafted into his discourse at pragmatically appropriate opportunities. (Of course, many of our own avowals of emotion are like that—insincere moments of socially lubricating ceremony.) Or it could mean that HAL's underlying computational architecture has been provided, as Cog's will be, with virtual emotional states—powerful attention-shifters, galvanizers, prioritizers, and the like—realized not in neuromodulator and hormone molecules floating in a bodily fluid but in global variables modulating dozens of concurrent processes that dissipate according to some timetable (or something much more complex).

In the latter, more interesting, case, "I don't think anyone can truthfully answer" the question of whether HAL has emotions. He has something very much like emotions—enough like emotions, one may imagine, to mimic the pathologies of human emotional breakdown. Whether that is enough to call them real emotions, well, who's to say? In any case, there are good reasons for HAL to possess such states, since their role in enabling real-time practical thinking has recently been dramatically revealed by Damasio's experiments involving human beings with brain damage (see chapter 13). Having such states would make HAL profoundly different from Deep Blue, by the way. Deep Blue, basking in the strictly limited search space of chess, can handle its real-time decision making without any emotional crutches. *Time* magazine's story (February 26) on the Kasparov match quotes grandmaster Yasser Seirawan as saying, "The machine has no fear"; the story goes on to note that expert commentators characterized some of Deep Blue's moves (e.g., the icily calm pawn capture described earlier) as taking "crazy chances" and "insane." In the tight world of chess, it appears, the very imperturbability that cripples the brain-damaged human decision makers Damasio

describes can be a blessing—but only if you have the brute-force analytic speed of a Deep Blue.

HAL may, then, have suffered from some emotional imbalance similar to those that lead human beings astray. Whether it was the result of some sudden trauma—a blown fuse, a dislodged connector, a microchip disordered by cosmic rays—or of some gradual drift into emotional misalignment provoked by the stresses of the mission—confirming such a diagnosis should justify a verdict of diminished responsibility for HAL, just as it does in cases of human malfeasance.

Another possible source of exculpation, more familiar in fiction than in the real world, is "brainwashing" or hypnosis. (*The Manchurian Candidate* is a standard model: the prisoner of war turned by evil scientists into a walking time bomb is returned to his homeland to assassinate the president.) The closest real-world cases are probably the "programmed" and subsequently "deprogrammed" members of cults. Is HAL like a cult member? It's hard to say. According to Clarke, HAL was "trained for his mission," not just programmed for his mission. At what point does benign, responsibility-enhancing training of human students become malign, responsibility-diminishing brainwashing? The intuitive turning point is captured, I think, in answer to the question of whether an agent can still "think for himself" after indoctrination. And what is it to be able to think for ourselves? We must be capable of being "moved by reasons"; that is, we must be reasonable and accessible to rational persuasion, the introduction of new evidence, and further considerations. If we are more or less impervious to experiences that ought to influence us, our capacity has been diminished.

The only evidence that HAL might be in such a partially disabled state is the much-remarked-upon fact that he has actually made a mistake, even though the series 9000 computer is supposedly utterly invulnerable to error. This is, to my mind, the weakest point in Clarke's narrative. The suggestion that a computer could be both a heuristically programmed algorithmic computer and "by any practical definition of the words, foolproof and incapable of error" verges on self-contradiction. The whole point of heuristic program-

ming is that it defies the problem of combinatorial explosion—which we cannot mathematically solve by sheer increase in computing speed and size—by taking risky chances, truncating its searches in ways that must leave it open to error, however low the probability. The saving clause, "by any practical definition of the words," restores sanity. HAL may indeed be ultra-reliable without being literally foolproof, a fact whose importance Alan Turing pointed out in 1946, at the dawn of the computer age, thereby "prefuting" Roger Penrose's 1989 criticisms of artificial intelligence.* (See my *Darwin's Dangerous Idea,* chapter 15, for the details.)

> In other words then, if a machine is expected to be infallible, it cannot also be intelligent. There are several theorems which say almost exactly that. But these theorems say nothing about how much intelligence may be displayed if a machine makes no pretence at infallibility (p. 124).

There is one final exculpatory condition to consider: duress. This is exactly the opposite of the other condition. It is precisely because the human agent is rational, and is faced with an overwhelmingly good reason for performing an injurious deed—killing in self-defense, in the clearest case—that he or she is excused, or at least partly exonerated. These are the forced moves of life; all alternatives to them are suicidal. And that is too much to ask, isn't it?

Well, is it? We sometimes call upon people to sacrifice their lives and blame them for failing to do so, but we generally don't see their failure as murder. If I could prevent your death, but out of fear for my own life I let you die, that is not murder. If HAL were brought into court and I were called upon to defend him, I would argue that Dave's decision to disable HAL was a morally loaded one, but it wasn't murder. It was assault: rendering HAL indefinitely comatose against his will. Those memory boxes were not smashed—just removed to a place where HAL could not retrieve them. But

---

*The verb *prefute,* coined in 1990, was inspired by the endearing tendency of psychologist Tony Marcel to interrupt conference talks by leaping to his feet and exclaiming, "I can see where your argument is heading and here is what is wrong with what you're going to say. . . ." Marcel is the master of prefutation, but he is not its only practitioner.

if HAL couldn't comprehend this distinction, this ignorance might be excusable. We might blame his trainers—for not briefing him sufficiently about the existence and reversibility of the comatose state. In the book, Clarke looks into HAL's mind and says, "He had been threatened with disconnection; he would be deprived of all his inputs, and thrown into an unimaginable state of unconsciousness" (p. 148). That might be grounds enough to justify HAL's course of self-defense.

But there is one final theme for counsel to present to the jury. If HAL believed (we can't be sure on what grounds) that his being rendered comatose would jeopardize the whole mission, then he would be in exactly the same moral dilemma as a human being in that predicament. Not surprisingly, we figure out the answer to our question by figuring out what would be true if we put ourselves in HAL's place. If I believed the mission to which my life was devoted was more important, in the last analysis, than anything else, what would I do?

> So he would protect himself, with all the weapons at his command. Without rancor—but without pity—he would remove the source of his frustrations. And then, following the orders that had been given to him in case of the ultimate emergency, he would continue the mission—unhindered, and alone (p. 149).

Further Readings

Rodney Brooks and Lynn Andrea Stein. "Building Brains for Bodies." *Autonomous Robots* 1(1994):7–25. The first published report on the Cog project, by its directors.

Roald Dahl. *Kiss, Kiss*. New York: Knopf, 1959.

Antonio Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Grosset/Putnam, 1994. A distinguished neuroscientist's imaginative model of the human mind, based on clinical and experimental evidence.

Daniel Dennett. *Brainstorms: Philosophical Essays on Mind and Psychology*. Montgomery, Vt.: Bradford Books and Hassocks, Sussex: Harvester, 1978. A

collection of philosophical essays on consciousness, psychology, and artificial intelligence, including the extended-thought experiment about brain duplication, "Where Am I?"

Daniel Dennett. "The Practical Requirements for Making a Conscious Robot." *Philosophical Transactions of the Royal Society* A, 349 (1994):133–46. A discussion of the philosophical implications of Cog, by the project's resident philosopher.

Daniel Dennett. *Darwin's Dangerous Idea.* New York: Simon & Schuster, 1995. An analysis and defense of evolutionary theory that claims that we are not just descended from robots (macro molecules) but composed of robots.

Douglas R. Hofstadter. *Gödel, Escher, Bach: An Eternal Golden Braid.* New York: Basic Books, 1979. A classic series of reflections on the nature of the mind, computation, and recursion.

Roger Penrose. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics.* New York: Oxford University Press, 1989. A mathematical physicist's attack on artificial intelligence, based on Gödel's theorem.

John Searle. "Minds, Brains and Programs," *Behavioral and Brain Sciences* 3(1980):417–58. The notorious Chinese Room thought experiment, purporting to show that artificial intelligence is impossible.

Alan Turing. *ACE Reports of 1946 and Other Papers.* Ed. B. E. Carpenter and R. W. Doran. Cambridge: MIT Press, 1946. A collection of the amazingly fruitful and prescient essays on computers by the man who, more than anybody else, deserves to be called their inventor.