Why did HAL commit murder? (Part 3 and Final)

By Daniel C. Dennett

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 programming 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.)

**The verb prefute, coined in 1990, was inspired by the endearing tendency of psychologist Tony Marcel to interrupt conference talks by leaping to this 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.
In other words, 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 pretense at infallibility.

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 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". 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.
Further Readings:


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.

Daniel C. Dennett is University Professor and Co-director, Center for Cognitive Studies, at Tufts University in Medford, Massachusetts, USA.

He is the author of many books, including *DARWIN'S DANGEROUS IDEA AND CONSCIOUSNESS EXPLAINED,* and over 400 scholarly articles in philosophy, cognitive science, and evolutionary theory.

His latest book is *INTUITION PUMPS AND OTHER TOOLS FOR THINKING.*