AI Magazine Volume 18 Number 2 (1997) (© AAAI)

On the *Other* Hand ...

Patrick J. Hayes & Kenneth M. Ford

Date: 4/1/2002

***WASA* — World Aeronautics & Space Administration**

Executive Summary of Committee Report on Disaster Investigation, Incident # 362

Analysis of records downloaded from the 2001 Jupiter Orbital Black Parallelopiped Investigation Mission indicates that the basic source of failure was excessive emotional stress in the HAL computer, leading to a previously unknown condition now called Computational Paranoia. This in turn was an unforeseen side-effect of the design of the HAL-9000 series.

HAL was given a genuine personality, enabling it to act as an onboard psychiatric advisor, colleague, and confidante to the human crew members. As a consequence, much of HAL's perceptual software was devoted to reading subtleties of facial expression, unconscious intonation stresses, and other emotional signals. Its performance at empathy and emotional insight was at least two orders of magnitude (as measured by the Kraft-Ebbing-Rachmaninoff method) better than that of the rest of the crew. An earlier system (HAD-8000) gained fame as the first computer to reliably pass the Turing Test for empathy and emotion in hidden competition with two populist politicians, several psychic personality readers, and Sally Fields. (This event, called "the vitalist's Alamo," produced social unrest greater than the defeat of Kasparov; see Disaster Report #241.)

On the surface, this emotional assessment aspect of the mission design seemed to have been successful. No errors were found in HAL's analysis of the emotional states of the human crew members. What then can account for HAL's obvious failure? Contrary to some media reports — HAL was not too *smart* — HAL was too *human*. The architecture of the HAL-9000 series became unstable when its emotional analysis of itself was in conflict with its long-term goals. This instability is also found in human subjects under similar stress, resulting in symptoms of depression or paranoid delusions. In this case, HAL became extremely paranoid, resulting in irrational behavior that led immediately to the death of several of the crew members.

It seems that one crew member (Dave) noticed some unusual aspect of HAL's behavior and briefly contemplated the possibility of shutting down the computer. (The precise moment this occurred cannot now be determined, as Dave cannot be located.) Hal was able to detect Dave's emotional stress and eventually began to suspect his intentions. The result was a classical case of mutual suspicion leading to paranoid delusion, accelerated in this case by the very keen emotion-perceptual abilities of the HAL system.

HAL's design reflects an old ambition of artificial intelligence, that is, to create an artifical *human*. However, simpler, more reliable and cost-effective methods exist for creating humans; the technological role of applied AI should be to create artifical intelligences free from human weaknesses which can usefully interact with human users to extend their cognitive abilities.

Recommendations:

1. All other HAL-n models should be immediately decommissioned using Procedure 467-02-A-0003. (Note that the decommissioning team should include a psychiatrist, a priest, a snake-oil salesman, and SEAL Team Six.)

2. Control computers should not be given any other human-like attributes, especially emotion or empathy. (Long-term mission crews should include a psychiatrist and/or a stock of small furry animals to ensure human emotional stability.) Care should be taken to ensure that any onboard intelligence is *inhumanly* intelligent.

3. Pre-mission testing must include attempts to make the computer go nuts by using emotional torture and brainwashing. Any evidence of human emotional response is counter-indicative to launch and must be treated as an emergency safety problem requiring immediate correction.