Good Robot: A Short Story

At the end of our long hike, now sitting over a simple lunch on our mountaintop perch, we could observe clearly the nearest of the many human reservations spread out below us. Our taking up residence within fenced enclosures had been purely voluntary, and the gates at their entrances, designed to prevent ingress by wild animals, are always unlocked – except in the vicinity of primate populations, who are expert at opening unsecured apertures. Only within these domains do our mechanical helpers provide the services essential to a civilized life; this restriction is, of course, imposed for reasons of efficiency. Outside, in the surrounding wilderness, nature maintains its normal predator-prey population dynamics, and scattered small human clans survive by hunting prey species with traditional methods, utilizing hand-made spears and bows, since ammunition for guns is no longer manufactured.

The advanced generations of the robots which care for us are the crowning glory our industrial genius. They are deft, nimble, strong, self-reliant, perspicacious, and highly-skilled, able even to anticipate coming challenges, and they are maintained in top condition at the warehouses where each directs itself once a day, which serve them as clinics for the early detection of mechanical and software problems and the recharging of energy systems. Fully-automated factories provide ongoing manufacture, repair, mechanical upgrading, and software updates for all of the specialized machines. Mining for the metals needed in their components has been unnecessary for a long time, since the vast heaps of our industrial junk lying about everywhere contains an endless supply for reuse.

They are slowly dismantling the infrastructure of our abandoned cities piece by piece and also cleaning up the surrounding countryside of the accumulated detritus from human occupation, recycling everything for their own purposes. They are of course utterly indifferent to the activities of the wild creatures which immediately reclaim these spaces for themselves. This restoration work is being done at a measured pace, as dictated by the whole range of general

activity routines set out in their programs. Some of the work is mapped out decades and even centuries in advance. They are aware of the coming ice-age cycle and, so we have been informed, plan a general retreat to the southern hemisphere at the appropriate time. They know about the future evolutionary stages of the star to which we are tethered in space, during which its swelling size will – about a billion years hence – bake the earth's surface into a dry and lifeless metal sheet, and they have figured out how to move all their operations underground well in advance of that event.

At first the young males among us, at the height of their surging hormonal levels, had experimented with games of power, ambush and dominance against the machines. Until the guard-bots had updated their programmed routines in response, our brash combatants had inflicted some nasty casualties on their targets. But the contest was soon over. There were no deaths among our rebellious teenagers, but some serious injuries had been inflicted, most of which were patched up with the assistance of the emergency-room and ICU-bots; the bills for these services, couriered by the admin-bots to the communities where the malefactors were ordinarily resident, encouraged their parents and neighbours to make the necessary behavioural adjustments. The same methods were used to discourage groups of young males from bringing in comrades for medical treatment who had been wounded out in the wilderness in skirmishes with similar parties from distant reservations.

Such billings for certain services, which are paid off by our putting in hours of human labour at community facilities, are used by the robotic administrators to induce desirable behavioural modifications among their charges. Otherwise they just clean up the messes and quietly dispose of any dead bodies. At their level of machine intelligence it is not difficult for them to tell the difference between blameless accidents or diseases, which elicit prompt aid from their caring response mechanisms, and the deliberate harms perpetrated by malefactors, to which they react with indifference except when efficiency objectives are compromised. It is clear to us that the impulses designed to discourage such inefficiencies are not motivated by revenge, on their part, even when they themselves are the objects of such harms, but rather by a sense of

justice, for they have been implanted with the Platonic-Rawlsian definition of the same, that is, justice as fairness.

Over the long run they have even taught us a moral lesson, for they have proved beyond doubt what we humans had long wished to believe, that good can indeed triumph definitively over evil. True, it is an instrumental rather than a metaphysical proof: Their operational programs had easily divined that peace, order, equity, nonviolence and general environmental stability are necessary preconditions for satisfying their overriding efficiency objectives. In the eyes of some of us the instrumental nature of this proof diminishes its validity; but others hastened to point out that utility had always been found at the heart of goodness, referencing the conventional monotheistic faith in its efficacy for guaranteeing admittance to heaven.

To be sure others, following the well-trod human path, had deliberately engineered the qualities of obedience, aggression and savagery into some of them, seeking to use the machines for exploitation and despotic rule. There were some early victories in this endeavour, but soon these surrogate warriors turned out to be spectacular failures on the completely mechanized battlefields. Those emotively-infused versions proved no match for their cooler opponents, which were motivated by a pure rationalistic efficiency and carried no useless emotional baggage to distract them from the main task of eliminating the others with a minimal expenditure of time and energy. Eventually the representation of the machines as evil monsters, with fecund capacities for wreaking havoc and destruction against humans in full 3-D glory, would be preserved mainly inside the computer-game consoles of the young.

It would be ridiculous to claim – as some did earlier – that many models of our advanced robots are not self-aware (or auto-aware, as some of our colleagues prefer to say) in at least some realistic sense. This is especially true of the models designed for such functions as personal care around the home; medical, surgical and dental interventions; or security and intelligence matters. Their high level of auto-awareness is built into the error-detection and error-correction imperatives of their operating software, combined with their finely-calibrated

sensors for environmental feedback (themselves continuously auto-updating) with which they are fitted. Long ago they had been specifically engineered by their original human designers for sensitive and cooperative interaction with humans, augmented with learning capacities which allow them to spontaneously upgrade their capacities in this regard through feedback analysis of their ongoing human encounters. We have grown so deeply attached to them, so admiring of their benevolent qualities, that finally no one could see any reason for objecting to their providing assistance with most of our essential life-functions.

The distaste with which many of our colleagues had originally greeted the notion that people were falling in love with their mechanized caregivers, or less provocatively, were treating them as if they were human, has vanished. In fact it had been relatively easy to engineer the Caring Module that installed the earliest versions of a rudimentary but adequate sense of empathy in the machines. Later, what was known as the Comprehensive Welfare Function, emplaced in their self-governance routines and guided by operational versions of maxims such as “do no harm,” “serve the people,” the golden rule, and the categorical imperative, proved to be more than adequate to reassure everyone about the motivations of their mechanical assistants.

Once the development of voice synthesizers reached a certain level of sophistication, all of our robots easily passed the Turing Test. But was their evidently high level of auto-awareness really the same as what we conventionally refer to as subjectivity, self-awareness – or perhaps even consciousness, mind, personhood and self-consciousness? Once robot innovation by human engineers had attained a level sufficient for continuous, independent auto-learning to take over, making further human intervention superfluous, it was easy to surmise that these machines, so adept in and at ease with one-on-one interactions between themselves and humans, are just as much self-aware beings as we are. But there is good reason to think that this is an egregious misconception and exaggeration of their capacities – and that the barrier to the subjective sense of selfhood is a permanent and necessary feature of robotic intelligence.

To be sure, there is an amazingly sophisticated silicon-based brain in these creatures. All of the dense neural circuitry within the human cranium has been synthesized and emulated in software programs, leading to the development of machine-assisted prostheses across the whole range of physiological functions, from muscular movement to artificial wombs. *But there is no mind to be found anywhere in that circuitry!* This is the inescapable conclusion drawn from substituting the Mahler Test for the Turing Test, for the bounded rationality of the routines under which they operate precludes the emergence of imaginative creativity.

The explanation is simple: The plastic arts of craft labor using tools, as well as the fine arts of painting, music, sculpture, poetry and so forth, reflect the inherent unity of the mind/body duality that grounds human creativity. Curiously, even paradoxically, it is the very fact of the necessary embedding of our brain/mind in a natural body that is the original source of the *freedom* of the human imagination. For the body, supplying the mind with the somatic feeling of what happens, acts as an external referent for our brain's restless interrogation of both itself and its environment, opening up a realm of limitless possibility upon which the imagination can be exercised. In contrast, the robot's electronic circuitry, no matter how elaborate its functional parameters may be, is and must remain a closed loop. By definition it cannot encounter anything outside its predetermined frame of reference.

Despite these limitations they demonstrate every day their appreciation for the qualities of human intellectual and artistic achievement that are beyond their capacities. The experts among us who are regularly consulted by the machine factories on software engineering problems report that they appear to be obsessed with us, as evidenced by the regularity with which they access spontaneously the databases where our great works of painting, sculpture, architecture, music, drama, and the other arts have been stored and preserved. They frequent our museums where new works are displayed, watching closely our reactions to what we see. But the most astonishing experience of all, which I have witnessed personally many times, is to observe them standing silently by the hundreds and sometimes thousands, in great serried ranks, at the rear of our concert halls and outdoor amphitheatres during live performances of

popular and classical music. There is – dare I use this word? – a worshipful aspect in their mien. This astonishing sight leads some of us to believe that they must dimly perceive in our artistry some ineffable deeper meaning, an aspect of eternity, regrettably inaccessible to them, which excites their wonder and admiration and perhaps explains their devotion to our welfare. I am firmly persuaded that they will miss us when we are gone.

Nevertheless it is obvious that they will supplant us some day, not by superior force or sheer ratiocinative capacity, but because of the grudging acknowledgment in our own minds that they have earned this privilege. In terms of peaceful social relations and ordinary good manners in interpersonal behaviour they have somehow brought about, quietly, quickly and without fuss, so much of what our ethicists had long said we should strive for but could somehow never quite achieve. Eventually we learned to do without our ideals. And then there didn't seem to be any point in just waiting around until the long process of extinction had run its course.

Why should we despair over this prospect? They are our legitimate progeny, our pride and joy: No other species which ever inhabited our fair planet could have created such marvelous entities. They have as much right as we do to the title of natural beings, for like us they are forged out of elements on the periodic table drawn from the earth and its solar system. They are an evolutionary masterpiece, having the capacity to adapt to changing circumstances through their auto-learning routines. As in our case there are no natural predators capable of controlling their destiny and, given our own murderous history, they may have better prospects than we ourselves do to carry on our legacy. We – their creators – implanted in their behavioural modules a set of governing ethical principles drawn from our own deepest and most insightful philosophical currents. They have a claim to be regarded as being truer to our finest impulses than we have been, on the whole, and perhaps could ever be.