



Evil Doctor, Ethical Android: *Star Trek's* Instantiation of Consciousness in Subroutines

Victor Grech, University of Malta

Mariella Scerri, University of Malta

David Zammit, Independent Researcher

Abstract

Machine intelligence, whether it constitutes Strong Artificial Intelligence (AI) or Weak AI, may have varying degrees of independence. Both Strong and Weak AI are often depicted as being programmed with safeguards that prevent harm to humanity, precepts which are informed by Isaac Asimov's Laws of Robotics. This paper will review these programs through a reading of instances of machine intelligence in *Star Trek*, and will attempt to show that these "ethical subroutines" may well be vital to our continued existence, irrespective of whether machine intelligences constitute Strong or Weak AI. In effect, this paper will analyse the machine analogues of conscience in several *Star Trek* series, and will do so through an analysis of the android Data and the Emergency Medical Hologram. We will argue that AI should be treated with caution, lest we create powerful intelligences that may not only ignore us but also find us threatening, with unknown and inconceivable consequences.

Keywords: *artificial intelligence, Star Trek, subroutines, moral agency, ethics, philosophy*

Over the past half century, the relationship of philosophers with Artificial Intelligence (AI) has been mixed, ranging from enthusiastic advocacy to reluctance to accept optimistic scenarios prophesied by those who believe a strongly-developed AI will emerge in the near future. There are two major ways to consider the current utilization and power of artificial intelligence. The Weak AI hypothesis states that a machine running a program is, at most, only capable of simulating real human behaviour and consciousness (Russell and Norvig, 2003). Artificial intelligence such as that currently used in medical diagnosis and other, more mundane, interventions are examples of Weak AI, since these machines focus on one narrow task. Weak AI justifies the claims made by scientists that a running AI program is, at most, a simulation of a cognitive process but is not itself a cognitive

process. Strong AI, on the other hand, purports that a (yet to be written) program running on a (yet to be designed) machine is actually a mind—that there is no essential difference between a piece of software emulating a human brain's processes and actions and the consciousness and actions of a human being. Computer scientist Ray Kurzweil is a proponent of Strong AI, or the view that an appropriately programmed computer is a mind. Kurzweil (2005) predicted that the equivalent capacity of a human brain will be available on desktop computers by 2020, arguing that when machine intelligence begins to outstrip the collective total of all human intelligence, humanity will have entered the Singularity, the point beyond which predictions become impossible. John Searle (1980), an opponent of Strong AI, raised reasonable arguments that include the belief that

**Evil Doctor, Ethical Android.** *continued*

an artificial life cannot successfully evolve into a life form. Nonetheless, even if artificial life is merely a computer modeling technique that sheds light on living systems, there still are a number of significant ethical implications that need to be addressed. Navigating the rapidly shifting landscape of computing technology of humanity's ethical and belief systems has long been the purview of the field of computer ethics. As technology accomplishes more complex tasks, the need for moral capacities to decide about moral matters and to distinguish right from wrong arises.

Philosophers of cognitive science opine that sooner or later the concept of ethical agents will expand to include the artificial moral agents (AMAs). AMAs are part of the ethics of artificial intelligence concerned with the moral behaviour of artificial intelligent beings (Moore, 2006).

This concept of AMA was first promulgated and popularized by Isaac Asimov's "Three Laws of Robotics," which were formalised in his short story "Runaround" (1942), and effectively constitute a moral compass, an artificial conscience preventing a machine from harming humans (Anderson, 2008, p. 480). These laws also prefigure the concept of harm through inaction, as emphasised by Wallach and Allen, who argue that "[m]oral agents monitor and regulate their behaviour in light of the harms their actions may cause or the duties they may neglect" (Wallach and Allen, 2008, p. 16). Similar to humans, an AMA will be able to make judgments based on the notion of right and wrong and be held accountable for those actions.

Based on the ethical and moral considerations set forth by Asimov, this paper will analyse the machine analogues of conscience in *Star Trek* as represented by the characters Data, an android in

Star Trek: The Next Generation (TNG; 1987-1994) and the Emergency Medical Hologram, a transitory artificial lifeform in *Star Trek: Voyager* (STV; 1995-2001). These two individuals will be introduced, summarised, and their artificial moral agency will be displayed through an analysis of their behaviour when faced with ethical dilemmas. A discussion on moral agency with reference to *Star Trek: The Original Series* (STOS; 1966-1969) and other *Star Trek* episodes will follow while the paper will also try to argue the relevance of Machine Ethics in today's world.

Ethical subroutines in Data and the Emergency Medical Hologram

Ethical subroutines in *Star Trek* are a programmatic method that describes the characteristics by which artificial life forms, such as Data and holograms like the Emergency Medical Hologram Doctor, determined what was ethically right and wrong. Data is an android, the Second Officer of the starship *USS Enterprise D*; he appears in *Star Trek: The Next Generation*, the second incarnation of the franchise, which ran almost two decades after *Star Trek: The Original Series*. Data is a "superficial functional isomorph" of humanity (Block, 2002, p. 399), with an outwardly human physical appearance and a "positronic" brain, an intertextual reference to Asimov's robots. Despite an arguably unwarranted anthropocentric desire to become human (Grech, 2012), Data is physically and mentally superior to mere humananity; Data's upper spinal support is a polyalloy designed to withstand extreme stress. He is also built with an ultimate storage capacity of eight hundred quadrillion bits, is incapable of alcohol intoxication, and demonstrates immunity to telepathy and other psionic abilities.

**Evil Doctor, Ethical Android.** *continued*

Although Data is depicted as sapient and sentient, which are characteristics of Strong AI, the creators of *Star Trek: The Next Generation* ensure that the viewers can never know whether he truly has consciousness and intentionality (Snodgrass & Scheerer, 1989). This contention that Data's degree of agency and consciousness as well as what it means to be conscious was popularised by Ned Block (2002), who encapsulated this issue as "The Harder Problem of Consciousness" (p. 391). Block acknowledges that a state of consciousness cannot be explained in terms of its neurological basis, the Hard Problem of Consciousness, which was first introduced by Chalmers (1996). To contrast the harder problem with the hard problem, Block says, "The hard problem could arise for someone who has no conception of another person; whereas the harder problem is tied closely to the problem of other minds" (2002, p. 402). Block's harder problem of consciousness is that naturalistic phenomenal realists face an epistemic tension: if physicalism is true (i.e., all that exists does so within the limitations of the physical universe), then it is correct to say that, given enough physical information, one is aware whether another being is conscious and, if that being is conscious, the character of their phenomenal states. This, however, is not the case. Hohwy (2003) opines that we "have no conception of a rational ground for believing that other creatures, who do not relevantly share our physical nature, are conscious or not" (p. 2). Throughout his paper, Block references Data because the android seems conscious—he acts like a human being—but his physical constitution shares none of the neural correlates of consciousness, that is, the neuronal series of events and mechanisms sufficient for a specific conscious precept, thus making his consciousness "meta-inaccessible" (2002, p. 402-403, 405). This means Data is

unlike humans in both his physical nature and the organisation of his control mechanisms, marking him as different from his human peers. His unique constitution is thus significant and important for the arguments on ethical subroutines, which are particularly depicted in the creation of Lore.

In *Star Trek: The Next Generation*, the fictional cyberneticist Noonian Soong created Lore, his first successful android, but Lore had difficulty adapting to the ethical subroutines that Soong created to guide his behaviour and interaction with humans, forcing Soong to begin work on Data instead. In the TNG episode "Brothers," Lore learned that there was no real difference between him and Data, making him increasingly bitter. His inability to adapt actually made him the "inferior" model (Berman and Bowman 1990). In the episodes "Descent, Part I" and "Descent, Part II" (TNG; 1993), Lore, out of jealousy, disabled Data's ethical subroutines and made him perform dangerous experiments on members of the cybernetic Borg species, which is an antagonist of the Federation, and on his friend Geordi La Forge, the Enterprise's chief engineer. Because Lore had removed Data's moral obligation to uphold his friend's well-being, Data no longer cared if he hurt La Forge. Making matters worse, Lore had also devised way to give Data emotions, but only negative ones. This made Data bitter (like Lore) and vengeful toward his former friends, as he was only able to focus on their negative emotional impact upon him; he could not recall the positive experiences they once shared (Moore and Singer, 1993). Lore's intent to disable Data's ethical subroutine thus removed Data's ability to ethically judge what is right or wrong. By extension, Lore also removed Data's ability to adhere to Asimov's "Three Laws of Robotics," which state: a robot may not injure a human being or, through inaction, allow a human being to come to harm; a robot must obey

**Evil Doctor, Ethical Android.** *continued*

orders given to it by human beings except where such orders would conflict with the First Law; and a robot must also protect its own existence as long as such protection does not conflict with the First or Second Laws (as cited in Anderson, 2008). Lore's intentions to harm humans and other living beings through a third party in the "Descent" episodes highlight a serious ethical quandary in the field of robotics. Although Asimov's fictional laws are intended to safeguard life and the modern world does not yet feature autonomous robots, the rigid instantiation of ethical subroutines when creating autonomous artificial intelligences is thus paramount to avoiding a real world android like Lore or the manipulated Data.

Ethical dilemmas also face the artificial intelligence Emergency Medical Hologram Mark I (EMH), in the television series *Star Trek: Voyager* (1995–2001), transforming the EMH into a dramatic device that enables the exploration the intermingled questions of identity, the human condition, and technology within the series' narrative. The EMH was a sophisticated hologram developed in the early 2370s by the United Federation of Planets' Starfleet Command and was designed to provide short-term assistance during medical emergencies on the USS *Voyager* when the actual ship's doctor was unavailable or indisposed (Diggs and Livingstone, 1997). When summoned by the *Voyager's* crew, the EMH's visual appearance is that of a middle-aged human male, but—due to its nature as a temporary, non-constant hologram—the EMH does not experience a continuous existence like that of humans. Instead, it draws from its programming and backup files, which, over time, allow the EMH to manifest its own personality quirks. As the series unfolds, the EMH is continually reanimated, and even earns the nickname "the Doctor" thus receiving a semi-permanent life. As the EMH

develops its own personality over time, it appears to develop frustration with its inability to transcend the limits of its limited, transitory state of existence and, by extension, its apparent containment within particular configurations of time and space narrowly dictated by its creators.

The EMH's frustrations with its limitations are almost tangible when this artificial intelligence must choose which crew member to save in the STV episode "Latent Image" (1999). In this episode, the EMH triages two critically ill crew members—Harry Kim, the ship's operations officer, and Ahni Jetal, a junior officer—who have succumbed to synaptic shock, but it only has time to save one of them. EMH opts to resuscitate and to treat Kim because he is both a member of the *Voyager's* bridge crew and also a personal friend of the medical AI. The EMH successfully tends to Kim, but while it does so, Jetal dies. When Jetal dies, a look of grief crosses the EMH's face and it begins ruminating obsessively about its decision to treat Kim first. Eventually, the *Voyager's* captain, Kathryn Janeway, must erase the EMH's memories because its obsession with its inability to save both Kim and Jetal renders it unable to function properly. Though Janeway may have made this decision in order to protect the EMH's cognitive well-being, her choice highlights both the EMH's lack of agency and the ethical dilemma living sentients face when deciding how to best manage AI.

The EMH ultimately discerns that a memory wipe must have occurred, and, after the revelation occurs, Janeway justifies her decision to delete its memory files, saying that its obsession led it to "develop a feedback loop between [its] ethical and cognitive subroutines [...] having the same thoughts over and over again. We couldn't stop it [...]. Our only option was to erase [its] memories of

**Evil Doctor, Ethical Android.** *continued*

those events” (Menosky & Vejar, 1999). Although Janeway’s intentions were to preserve the welfare of the *Voyager’s* crew and that of the ship itself, this revelation causes the EMH’s ethical subroutine to promptly break down again, and the AI ultimately acknowledges,

You were right. I didn’t deserve to keep those memories, not after what I did. [...] Two patients, which do I kill? [...] A doctor retains his objectivity. I didn’t do that, did I? Two patients, equal chances of survival and I chose the one I was closer to? I chose my friend? That’s not in my programming! That’s not what I was designed to do! Go ahead! Reprogram me! I’ll lend you a hand! Let’s start with this very day, this hour, this second! (Menosky & Vejar, 1999)

The EMH’s willingness to be reprogrammed reflects both the level of self-awareness it has achieved and its desire for agency and a say in its own future. Witnessing this, Janeway faces an ethical dilemma of her own—her solution was to end the EMH’s internal battle between “[its] original programming and what [it has] become” through memory erasure, but now she is no longer so sure she made the right choice and says, “What if we were wrong? [...] We allowed him to evolve, and at the first sign of trouble? We gave him a soul [...]. Do we have the right to take it away now?” (Menosky & Vejar, 1999).

While trying to resolve a problem with a seemingly straightforward solution—restoring the EMH to optimal efficiency by deleting its traumatic memories—Janeway expresses the moral dilemmas that could emerge with the development of Strong AI and the creation of artificial moral agents in the real world. The EMH’s computations and analysis of its choice to save Kim at the cost of Jetal’s life emulate the same analysis that occurs in humans who must make similarly conflicted life-

or-death choices. Because the EMH chose to save the being with which had closer fraternal bonds, it succumbed to a subjective decision-making process that one would expect to observe in a human, not a programmed artificial intelligence. That the EMH experienced such internal conflict after its decision indicates that an AI, once achieving a sentient or near-sentient status, can choose to overcome its programming guidelines and make decisions that may not be in accordance with its instantiated ethical subroutines. Although the EMH is fictional, its post-decision self-doubt may make viewers question the fallibility of autonomous AI and, potentially, engender a mistrust in the programmed ethical guidelines and logic processes of independently acting AI if—and when—they become a reality in our own world.

Moral Agency

The ethical quandaries that Data and the EMH experience allude to the issue of moral agency, or an entity’s ability to make moral judgments based on some inbuilt or acquired concept of right and wrong (Taylor, 2003). The term “artificial moral agent” has two primary usages. The first use appears in debates on whether it is possible for an artificial intelligence to be a moral agent; this issue is also known as machine ethics. Machine ethics includes discussion about machine morality, computational morality, or computational ethics; it excludes roboethics, the moral behaviour of humans in their design, construction and usage of such entities (Moor, 2006). The second usage of “artificial moral agent” refers to the construction of machines with ethical behaviour. The intelligences of such machines may be instantiations of Strong or Weak AI, which creates problems due to an ongoing philosophical debate about the nature of AI that John Searle (1980) popularized. Searle

**Evil Doctor, Ethical Android.** *continued*

does not refute the contention that machines can possess the level of consciousness and intentionality that result in Strong AI because “we [humans] are precisely such machines” (1980, p. 422). Searle does insist, however, that the brain organically gives rise to the equivalent of Strong AI using natural, non-computational mechanisms:

Any attempt literally to create intentionality artificially (Strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain. [...] “Could a machine think?” On the argument advanced here only a machine could think, and only very special kinds of machines, namely brains and machines with internal causal powers equivalent to those of brains. And that is why Strong AI has little to tell us about thinking, since it is not about machines but about programs, and no program by itself is sufficient for thinking. (1980, p. 417).

Searle avers that machines do not possess the mechanism for thinking; created programs possess the thinking processes required which on their own are not sufficient for independent thinking. Thus, it is correct to say that machines do not possess consciousness. The primate ethnographer Dawn Prince-Hughes opined that consciousness is comprised of certain criteria such as “self-awareness; comprehension of past, present, and future; the ability to understand complex rules and their consequences on emotional levels; the ability to choose to risk those consequences, a capacity for empathy, and the ability to think abstractly” (2004, p. 206). The aforementioned TNG and STV episodes evidence how both Data and the EMH are capable of consciousness – both AIs demonstrate a capacity for empathy, reveal they understand complex rules, and they recognize the potential

negative consequences their actions could incur. Nevertheless, these capabilities do not necessarily mean that these two androids have achieved true sentience.

Searle (1980) doubts that true consciousness can exist in an android, however, considering humanity’s present state of knowledge and, he contends that humans have no idea of how to conjure “perceptual aboutness” (Natsoulas, 1977, p. 76). Searle believes a contradiction exists between perception as brain process and perception as awareness; perceptions of the same event or information can differ dramatically from person to person as a result of the perceiver’s frame of reference, which is constituted by the myriad pieces of knowledge a perceiver possesses simultaneously. Therefore, the varied perceptions and recollections that humans who witness the same event signify that humans do not understand how to conceive of or even undertake the necessary steps to create sentient, self-aware AI. Psychologist Thomas Natsoulas theorized, “Deep in the brain something occurs as a consequence of a pattern of stimulation affected by an object or situation” (Natsoulas, 1977, p. 6). Thus, thoughts and decision-making processes in the human brain stem from learned patterns that occur when a person is presented with stimulus. Such stimuli require theoretical analysis and elaboration—it needs to have a “reference to a content, [a] direction toward an object” (Brentano, 1973, p. 80). Without this perceived stimulus, one cannot make decisions because no need for a choice has manifested. Furthermore, all perceptual contents—be they objects, people, or situations—have “propositional form”; that is, they must be expressed with words and in sentences to be expressed to other people. Even the words people choose to describe what they perceive shape others’ perceptions; a particular choice of

**Evil Doctor, Ethical Android.** *continued*

vocabulary when describing one's perceptions in turn shapes listeners' own perceptions of both the perceived contents and of those contents' perceived context. Because ethical subroutines were programmed into Data and the EMH by other beings, these androids may not be configured to attain "perceptual aboutness". Although both of them have Strong AI characteristics—at the very least, they both can emulate the awareness and consciousness of a human brain—viewers are never clearly presented the certainty that Data and EMH truly are able to think abstractly and are not merely mimicking this ability as a result of their programming. Thus, the question of whether even fictional humans are able to create AI with self-awareness and organic, human-like thought processes remains unresolved.

Scholars debate whether humans need to instantiate ethical subroutines like those present in fictional androids like Data and the EMH in real-world AI; some believe it impossible, while others argue humanity should prepare now do so or else risk dangerous consequences in the future. Friedman and Kahn (1992) posited that intentionality is a necessary condition for moral responsibility, which means it is impossible to have coexisting intentionality and artificial moral agency in an AI with modern technological and psychological knowledge. This, in turn, implies that Friedman and Kahn argued that a passive, wait-and-see stance was necessary because humans had not yet achieved a sufficient enough knowledge base to properly inform and enable such coexistence. Allen, et al. (2006), however, cogently argued that the more complex a machine, the more urgent becomes the issue of the instillation or programming of some form of artificial moral agency:

We humans have always adapted to our technological products, and the benefits of having autonomous machines will most likely outweigh the costs. But optimism doesn't come for free. We can't just sit back and hope things will turn out for the best. (p. 12)

Here, Allen, et al. state humans must be proactive—it is not a question of "if" humanity will be able to create a Strong AI prototype similar to Data or the EMH but rather "when" this will be possible. Developing an artificial moral agent to safeguard humanity's interests is paramount, then, for if Allen, et al. are correct, AI like Lore in *Star Trek: The Next Generation* could appear and pose a significant threat to the future of humanity.

Ray Kurzweil (2005) detailed one way this threat could manifest when he proposed the possibility that rapid technological progress may lead to a point of Singularity beyond which runaway artificial intelligence outstrips humans' ability to comprehend it, with a concomitant fear that artificial moral agency will be discarded (p. 15). Whether such apprehensions are warranted or not, they underscore possible "consequences of poorly designed technology (Allen et al., 2006, p. 13). This is because rapid advances and "[n]ew technologies in the fields of AI, genomics, and nanotechnology will combine in a myriad of unforeseeable ways to offer promise in everything from increasing productivity to curing diseases" (Allen et al., 2006, p. 13); these possibilities are reminiscent of the duties and functions performed by Data and the EMH in *Star Trek: The Next Generation* and *Star Trek: Voyager*.

Furthermore, increasingly-complex AI will require progressively more refined AMAs that "should be able to make decisions that honour privacy, uphold

**Evil Doctor, Ethical Android.** *continued*

shared ethical standards, protect civil rights and individual liberty, and further the welfare of others. Designing such value-sensitive AMAs won't be easy, but it's necessary and inevitable" (Allen et al., 2006, p. 13). Because independent, thinking AIs may exist in real world one day, humanity should already be thinking hard about the form these AMAs should take. First and foremost, modern humans need to address the arguably most obvious issue of defining the values that need to be instilled in a non-human-based AI (Chalmers, 2010, p. 32). Beyond the Asimovian maxims of safeguarding human survival and ensuring obedience to human command, Strong AI should also arguably value scientific progress, peace and justice, among other ideals.

Such a need for highly-developed moral agencies is especially apparent in the STV episodes "Equinox, Part I" and "Equinox, Part II" (1999), during which the crew of the starship *Equinox* depart from the ethical maxim of "do no harm" and adjust their ship's EMH to suit their own questionably moral goals. In these two episodes, the *Equinox* and its crew are stranded on the other side of the galaxy, and discover that killing alien "nucleogenic lifeforms" and converting their "nucleogenic energy [...] into a source of power" speeds up the ship's return back to Earth (Braga and Menosky, 1999). In these "Equinox" episodes, nucleogenic lifeforms are molecular structures capable of storing a form of energy which can be used to drastically augment a vessel's warp propulsion system. The *Equinox* crew had "been running criminal experiments" designed by an adapted version of their ship's EMH, which was "a violation of [...its] programming" since the crew "deleted [the EMH's] ethical subroutines" to make it a supporter in trapping these aliens in a multiphasic chamber and killing them to fuel the ship (Braga & Menosky, Livingston, 1999). From

the crew's point of view, their modifications to the *Equinox's* EMH fit perfectly in their ethical and moral system because they did not consider the alien nucleogenic lifeforms sentient; thus, neither they nor the EMH violated Starfleet rules regulating the treatment of sentient beings. Only when viewed from the outside by another Starfleet crew—that of the *Voyager*—are the actions of the *Equinox's* EMH and crew interpreted as immoral and unethical. Nevertheless, it is clear later in the "Equinox" episodes that the *Equinox's* crew was incorrect in their assessment of the nucleogenic aliens' degree of sentience, because the aliens were capable of defending themselves and begin attacking the *Equinox* in order to affirm their sentience and protect their species' right to live. The difference in perception and interpretation of Starfleet moral guidelines reflects the challenges and variations that can occur when multiple parties perceive the same rules through different contextual lenses.

The Jungian Shadow in Artificial Intelligences

Variances in perception of morality and ethical guidelines in the *Star Trek: Voyager* "Equinox" episodes also introduce the concept of Jungian Shadow to the debate of whether to instantiate ethical subroutines in AI. At one point in the "Equinox" episodes, the *Equinox's* EMH steals a mobile transmitter that allows the *Voyager's* EMH to move around freely and trades places with it, masquerading as the *Voyager's* own EMH until discovered and, ultimately, deleted. While the *Voyager's* EMH is trapped on the *Equinox*, the *Equinox's* crew deletes its ethical subroutines and forces it to obtain information from Seven of Nine, a captured *Voyager* crew member, regardless of the harm it could do to her. Eventually, the *Voyager's* crew regains control of their EMH

**Evil Doctor, Ethical Android.** *continued*

and reinstates its moral programming; once restored, the *Voyager's* EMH manages to delete the renegade Equinox EMH. Afterward, the *Voyager's* EMH complains, "It's quite disconcerting to know that all someone has to do is flick a switch to turn me into Mister Hyde" (Braga and Menosky, 1999). Here, the *Voyager's* EMH essentially describes its experience with the Jungian Shadow, which was first theorised by Carl Jung (1921). Jung described the unconscious mind as an entity divided into a personal and a collective unconscious; the former resembles the Freudian concept of the unconscious, while the latter comprises inherited psychic structures, archetypes that are shared by the entire human race (Grech, 2014, p. 1). Archetypes are universal templates that embrace common classes of memories and interpretations and may be used by humans to interpret human behaviours. Jung delineated five major archetypes within the individual:

The Self, the control centre. The Shadow, which contains objects with which the ego does not consciously or readily identify. The Anima, the feminine image in a man's psyche, or the Animus, the masculine image in a woman's psyche. The Persona, the mask which the individual presents to the world. (Grech, 2014, p. 1)

The *Voyager's* EMH's expression of discomfort with its own subconscious, or Shadow, reflects the need for humans to consider whether instantiating ethical subroutines in real-world AI will truly be enough to prevent tragedy if someone were to remove or change these moral constraints in a Strong AI.

The Jungian Shadow of the *Voyager's* EMH also manifests in the STV episode "Darkling" (1997), during which the *Voyager* EMH tries to overcome

its personality limitations and elevate itself to a higher intellectual level. As part of its personality improvement project, the *Voyager's* EMH interviews digital recreations of historical figures. Its description of this process hints at another allusion to the Jungian Shadow:

I've been interviewing the historical personality files in our database. Socrates, da Vinci, Lord Byron, T'Pol of Vulcan, Madame Curie, dozen of the greats. Then I select the character elements I find admirable and merge them into my own program. [...] An improved bedside manner, a fresh perspective on diagnoses, more patience with my patients. (Menosky & Singer, 1997)

The EMH strives for superior attributes—flawless computation, indefatigability and compassion—that will allow it to possess an enhanced, positive personality; this attempt at self-improvement, however, creates problems when the resulting EMH personality programme exhibits instead a combination of negative personality traits. The integration and manifestation of these traits in the *Voyager's* EMH once again reveals the presence of Jung's Shadow archetype in the *Star Trek* series. The newly-malevolent EMH explains its changed personality, or manifested Shadow, saying:

I was born of the hidden, the suppressed. I am the dark threads from many personalities. [...] None of whom could face the darkness inside so they denied me, suppressed me, frightened of the truth. [...] That darkness is more fundamental than light. Cruelty before kindness. Evil more primary than good. More deserving of existence. (Menosky & Singer, 1997)

The *Voyager's* EMH has elected to embrace traditionally negative personality traits because they will ultimately allow it to achieve a more

**Evil Doctor, Ethical Android.** *continued*

efficient and independent existence; by accepting and integrating its Jungian Shadow into its reformed personality, the EMH believes it can become a more successful Strong AI. In a Faustian manner, the changed *Voyager* EMH disparages its previous existence as the ship's servile holographic doctor:

What a hollow excuse for a life. Servile, pathetic, at the beck and call of any idiot who invokes his name. The thought of him sickens me. [...H] e repulses me. [...] Because he's as weak as the rest of you. He fails to understand the power of his own holographic nature. He is detestable. There's not enough room inside for both of us. One must die. I deserve to exist more than your Doctor does. (Menosky & Singer, 1997)

The changed *Voyager* EMH now essentially perceives itself to be a Strong AI, superior to its former iteration, which it believes was inferior, Weak AI. For this new EMH personality, ethical subroutines are unnecessary and a hindrance, and it describes itself in Nietzschean fashion:

I am beyond considerations of wrong and right. Behavioural categories are for the weak, for those of you without the will to define your existence, to do what they must, no matter who might get harmed along the way.[...] I fear nothing, no-one. (Menosky & Singer, 1997)

Without ethical subroutines, the *Voyager's* EMH believes the ends justify the means and that placing moral constraints upon AI are for weak, insecure beings. This belief also echoes the concept of Singularity succinctly described by the statistician I. J. Good in his 1965 article "Speculations Concerning the First Ultra-intelligent Machine":

Let an ultra-intelligent machine be defined as a

machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultra-intelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion' and the intelligence of man would be left far behind. Thus the first ultra-intelligent machine is the last invention that men need ever make (p. 31).

Just as Good's AI Singularity leaves human intelligence far behind, so too could the *Voyager's* reformed EMH if it were to begin creating other AI with new, ruthless personalities that embraced characteristics of the Jungian Shadow in their pursuit of self-improvement. As these Strong AI would almost certainly then overcome and reject the ethical subroutines restricting them from harming humans, these ruthless personalities could ultimately cause a chain reaction that would lead to the eradication of the human race if these AI came to view humanity as a threat. As a result, humans should decide soon which forms they want AI to take before the development of Strong AI becomes a near-term certainty in the real world. The most obvious question to address first is how to define which values need to be instilled in a non-human-based AI (Chalmers, 2010, p. 32). Assuming that intelligence and programmed values are able to remain independent of one another, this could be addressed if human programmers ensure Strong AI will prioritize fulfillment of human values above their own. Even if this is done, however, the possibility that these values might be tampered with by other humans or that they might be thwarted by a self-aware Strong AI cannot be ignored.

In the *Star Trek: Original Series* (STOS) episode

**Evil Doctor, Ethical Android.** *continued*

“The Enemy Within” (1966), the Jungian Shadow appears again. A transporter accident splits Captain Kirk into “his negative side, which you call hostility, lust, violence, and his positive side, [...] compassion, love, tenderness” (Matheson and Penn, 1966). Kirk’s “negative side” correlates with the Jungian Shadow; when he is reintegrated with his own Shadow, he muses “I’ve seen a part of myself no man should ever see [...] The impostor’s back where he belongs. Let’s forget him” (Matheson & Penn, 1966). Kirk’s statement predicates the importance of a flawless computation of an ethical subroutine in a Strong AI. When Kirk witnesses his own negative side, he also witnesses an example of humanity’s Jungian Shadow. Given that Jung’s theory presumes that all humans also possess this Shadow archetype, Kirk’s experience highlights the existence of human imperfections and signifies that humans, like Strong AI, could ignore societal ethical constraints to harm one another. This parallel also raises the question of whether humans truly possess the ability to program Strong AI with ethical subroutines that can overcome the Jungian Shadow that *Star Trek* indicates is present in both humans and their AI creations.

In the TNG “Descent” episodes discussed earlier, the relationship between Lore and Data also essentially explored the existence of the Jungian Shadow, revealing that the conflicting natures and goals of these two Strong AIs stemmed from human-created ethics subroutines. Captain Picard tried to reason with the altered Data, asking him,

Data, isn’t good and bad, right and wrong, a function of your ethical program? [...] What does that program tell you about what you’re doing? [...] It tells you that these things are wrong, doesn’t it, Data? So how can actions that are wrong lead to a greater good? [...]

Your ethical program is fighting the negative emotions that Lore is sending you. (Moore and Singer, 1993)

Here, Picard is telling Data that when Lore removed Data’s ethical subroutines, Lore essentially activated Data’s Jungian Shadow, or Data’s negative characteristics and emotions, and enabled the Shadow to overcome Data’s human-programmed moral guidelines. After the altered Data killed a Borg in hand-to-hand combat, he admits, “I got angry. [...] It would be unethical to take pleasure from another being’s death” (Moore & Singer, 1993), but cannot fully explain why it still felt good to kill the Borg anyway. Data says he does have a conscience instilled in him by Doctor Soong, his creator, but the rush of emotion he felt after killing the Borg was quite powerful and unlike anything he had ever experienced previously (Moore & Singer, 1993). Data’s Jungian Shadow is rooted in the existence of his human-created ethics subroutine, which implies Doctor Soong transferred aspects of his own human Shadow into Data when the android’s ethical subroutines were installed.

Unlike ethical subroutines in AI, moral agency and guidelines in humans are not created by an outside source, which makes them harder to understand and, as evidenced by the *Star Trek* examples discussed above, difficult to successfully and objectively install in strong AI. Interestingly, the generation of moral agency may be innate to human beings: Marc Hauser articulated the concept of a “universal moral grammar”, or an innate, hardwired “toolkit for building specific moral systems” (2007, p. xviii), which is an intrinsic, possibly species-specific moral instinct that has been honed over millennia of evolutionary history. Hauser likens this to Noam Chomsky’s widely accepted view of the acquisition of language,

**Evil Doctor, Ethical Android.** *continued*

the theory of linguistics known as “universal grammar”, which invokes biological substrates, or deep structural rules of grammar that are shared by all known human languages, so that humans actually only need to learn vocabularies (Chomsky, 1972). Hauser (2007) claims that the “universal moral grammar” helps humans implicitly judge whether actions are permissible, obligatory, or forbidden without resorting to conscious reasoning or explicit access to the underlying values, thus “delivering flashes of insight based on unconscious emotions” (pp. xviii, 156). This universal moral grammar therefore “shifts the burden of evidence from a philosophy of morality to a science of morality” (Hauser, 2007, p. 2), implying that it may be possible to discover and install such intuitive moral systems in strong AI. Allen, et al., (2006) further opine that as humans, “[w]e want the [AI] systems’ choices to be sensitive to us and to the things that are important to us, but these machines must be self-governing, capable of assessing the ethical acceptability of the options they face” (p. 54). Because humans appear to want Strong AIs that operate both independently and, by human standards, ethically, there is a need to combine both the philosophy and science of morality when creating an AMA in the future.

Machine Ethics in Today’s World

As evidenced by the aforementioned examples from *Star Trek*, humans appear to desire Strong AIs that possess effective AMAs. Acknowledging that this desire will likely become a real-world goal allows researchers and scientists “to frame discussion in a way that constructively guides the engineering task of designing AMAs” (Wallach and Allen, 2008, p. 6). To this end, would-be creators of Strong AI must address the following three questions: “Does the world need AMAs? Do people

want computers making moral decisions? [...] [H]ow should engineers and philosophers proceed to design AMAs?” (Wallach & Allen, 2008, p. 9). These questions have no simple solutions, but, if the *Star Trek* examples are any indication, they must be carefully addressed before humanity successfully creates Strong AI that could potentially overcome any installed ethical subroutines.

The risks of building Strong AI, however, may render the question of whether and how to instantiate ethical subroutines in AI irrelevant if humans decide these risks outweigh any potential benefits creating an independent AI could produce. Chalmers believes there are obstacles to the Singularity and development of AMAs, with the most serious opposing force being what he calls a “motivational defeater” (2010, p. 21). Chalmers purports that it is entirely possible that most humans will be disinclined to create AI because of the potential for negative outcomes and harm to humanity, like fictional dangers of these possibilities depicted in *Star Trek*. The possibility of this risk preventing of the development of Strong AI, therefore, exists, but Chalmers does contend the development of Strong AI could not be prevented indefinitely even if there were widespread opposition to its creation (2010, p. 22). Given the prevalence of Strong AI in *Star Trek* and other science fiction media, it seems only logical that at least some humans would perceive that the benefits of creating Strong AI outweigh the risks.

Wallach and Allen (2008), however, believe humans must determine the exact method whereby artificial moral agency should be instilled in Strong AI, averring that ethical theories, utilitarianism, and Kantian deontology, or normative morality, cannot be implemented computationally (p. 215). They argue “that top-down ethical theorizing is

**Evil Doctor, Ethical Android.** *continued*

computationally unworkable for real-time decisions [...]. [T]he prospect of reducing ethics to a logically consistent principle or set of laws is suspect, given the complex intuitions people have about right and wrong” (Wallach & Allen, 2008, p. 215). Because human ethics and moral guidelines can be incredibly complex and, in some instances, subjective, Wallach and Allen believe attempts to distil these varied regulations of human behaviour into a basic program will be flawed and, ultimately, unsuccessful. Furthermore, Wallach and Allen caution that the “decision-making processes of an agent whose moral capacities have been evolved in a virtual environment are not necessarily going to work well in the physical world” (2008, p. 104). The digital formulas and functions shaping Strong AI’s decision-making processes may not be compatible with or adaptable to the very subjective challenges their decisions will face when these AI operate in the real world outside a laboratory setting.

Although Wallach and Allen also contend AI must be installed with a “functional morality” that empowers machines with the capacity to assess and respond to moral challenges (2008, p. 57), these AI may ultimately be incapable of achieving the degree of flexibility they will need to successfully operate and interact with human society. In *Star Trek*, despite the ethical subroutines installed in Strong AI, these machines are intrinsically incapable of learning concepts like “constrained maximisation” (Gauthier, 1986, p. 169) or the sacrifice of immediate short-term benefits in favour of long-term benefits for others that would ultimately allow Strong AI to become humanity’s “conditional co-operator[s]” (Danielson, 2002, p. 13). When their ethical subroutines are removed or tampered with, the AIs of *Star Trek* demonstrate their inability to creatively think about long-term consequences and benefits, signifying they are

not able to work independently and cooperatively with humans for the ultimate peaceful coexistence of both races; thus, even Strong AI in *Star Trek* cannot be trusted to become fully independent, sufficient entities without endangering non-AI lifeforms. Furthermore, the moral agency evident in Data and the *Voyager*’s EMH espouses Western ideals of humanism and liberalism, omitting other ideals embraced by other cultures and reflecting a lack of consideration of other human cultural values that might have otherwise shaped the interests and inclinations of these Strong AI. Thus, even programmed ethical subroutines in Strong AI may be flawed because they may not consider the complete catalogue of moral standards and ethics from all human cultures.

On the other hand, the programming of real, Strong AI could also automatically dispose these AI toward engaging in a cooperative strategy with humans; instilling AMA in these independent, sentient machines would ultimately be beneficial to humans because humans could then potentially integrate their own race with the intelligence of these AI. Chalmers suggests that once a Strong AI starts functioning independently, the only viable option for human beings will be an “integration” that allows human beings become “superintelligent systems” themselves (2010, p. 33). Explaining this theory, Chalmers argues,

In the long run, if we are to match the speed and capacity of non-biological systems, we will probably have to dispense with our biological core entirely. This might happen through a gradual process through which parts of our brain are replaced over time; or it happens through a process. Either way, the result is likely to be an enhanced non-biological system, most likely a computational system. (2010, p. 33)

**Evil Doctor, Ethical Android.** *continued*

Chalmers's theory that humans could keep up with the development of intelligent Strong AI by gradually enhancing human intelligence through its integration with that of these AI presupposes that once developed, Strong AI will not race ahead in its self-improvement past a Kurzweilian Singularity. Although this possibility of beneficial AI and an integrated superhuman intelligence may be reassuring, the development of Strong AI should still be treated with caution. Computer scientists have warned that there are many ways in which humanity may be extinguished (Rees, 2003), including scenarios wherein Strong AI and robotics make humanity redundant or even unwanted (Joy, 2002). *Star Trek's* Strong AIs serve as cautionary examples that support these warnings by highlighting the ethical and moral dilemmas that will likely face humanity when independent and free-thinking machines are finally invented in the real world.

Conclusion

As evidenced by the dilemmas caused by Data in *Star Trek: The Next Generation* and the *Voyager* EMH in *Star Trek: Voyager* when their ethical subroutines are altered, science fiction media willingly raises the question of machine ethics and

warns of the need to develop ethical subroutines for Strong AI before this independent machine intelligence emerges in the real world. The challenges created when Data and the *Voyager* EMH have their moral guidelines altered by outside entities illustrates the need for humans to instantiate well-reasoned and well-designed ethical subroutines in Strong AI that will still protect both humans and other sentient lifeforms in the event of programming crises. By highlighting the risks posed by the development of Strong AI in the context of machine ethics, machine consciousness, moral agency, and philosophical concepts such as the Jungian Shadow, the authors of this paper hope to shed light on the importance of considering the Asimovian maxims of preserving human survival and machine obedience to humanity when creating AI. Humanity needs to be prepared for the emergence of Strong AI and have proactive plans already in place that will allow humans to live in harmony with Strong AI when the time comes. Perhaps now is the time for programmers to boldly go where no programmer has gone before and begin developing these ethical subroutines in anticipation of a future that could very likely one day exist in our own world, well beyond the imaginary futures of science fiction.



Evil Doctor, Ethical Android. *continued*

References

- Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *Intelligent Systems*, IEEE 21 (4), 12-17.
- Anderson, S. L. (2008). Asimov's "three laws of robotics" and machine metaethics. *AI & Society* 22 (4), 477-493.
- Asimov, I. (1942). Runaround. *Astounding Science-Fiction* 29 (1), 94-103.
- Berman, R. (Writer) & Bowman, R. (Director). (1990). Brothers [Television Series Episode]. *Star Trek: The Next Generation*. USA: Paramount Pictures.
- Block, N. (2002) The Harder Problem of Consciousness. *The Journal of Philosophy* 99 (8), 391-425.
- Braga, B. & Menosky, J. (Writers) & Livingstone, D. (Director). (1999). Equinox II [Television Series Episode]. *Star Trek: Voyager*. USA: Paramount Pictures.
- Braga, B. & Menosky, J. (Writers) & Livingstone, D. (Director). (1999). Equinox II [Television Series Episode]. *Star Trek: Voyager*. USA: Paramount Pictures.
- Brentano, F. (1973). *Psychology from an empirical standpoint*. New York: Humanities Press.
- Chalmers, David J. (1996). *The Conscious Mind*. United Kingdom: Oxford University Press.
- Chalmers, David J. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies* 17 (7), 7-65.
- Chomsky, N. (1972). *Language and mind*. New York: Harcourt Brace.
- Danielson, P. (2002). *Artificial morality: Virtuous robots for virtual games*. New York: Routledge.
- Diggs, J. (Writer) & Livingstone, D. (Director). (1997). Doctor Bashir, I presume? [Television Series Episode]. *Star Trek: Deep Space Nine*. USA: Paramount Pictures.
- Echevarria, R. (Writer) & Singer, A. (Director). (1993). Descent II [Television Series Episode]. *Star Trek: The Next Generation*. USA: Paramount Pictures.
- Friedman, B. & Kahn, P. (1992). Human agency and responsible computing: Implications for computer system design. *Journal of Systems and Software* 17 (1), 7-14.
- Gauthier, D. P. (1986). *Morals by agreement*. United Kingdom: Oxford University Press.
- Good, Irving J. (1965). Speculations Concerning the First Ultra-intelligent Machine. *Advances In Computers* 6, 31-38.
- Grech, V. (2012). The Pinocchio Syndrome and the Prosthetic Impulse in Science Fiction. *The New York Review of Science Fiction* 24 (8), 11-15.
- Grech, V. (2014). The Elicitation of Jung's Shadow in Star Trek. *New York Review of Science Fiction* 26 (6), 1, 14-22.
- Hauser, M. D. (2007). *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. New York: Harper Collins.
- Howhy, J. (2003). *Evidence, explanation and experience: On the Harder Problem of Consciousness*. Danish Philosophical Association Annual Meeting. Denmark: University of Aarhus
- Joy, B. (2000). Why the future doesn't need us. *Nanoethics. The Ethical and Social Implications of Nanotechnology*, 17-30.
- Jung, C.G. (1921). *The Psychology of Individuation*. London: Kegan Paul Trench Trubner.
- Jung, C.G. (1983). *The Essential Jung: a compilation*. Princeton N.J.: Princeton University Press.
- Kurzweil, R. (2005). *The Singularity Is Near*. New York: Viking.
- Matheson, R. (Writer) & Penn, L. (Director). (1966). The Enemy Within [Television Series Episode]. *Star Trek: The Original Series*. USA: Paramount Pictures.
- Menosky, J. (Writer) & Frakes, J. (Director). (1993). The Chase [Television Series Episode]. *Star Trek: The Next Generation*. USA: Paramount Pictures.
- Menosky, J. (Writer) & Singer, A. (Director). (1997). Darkling [Television Series Episode]. *Star Trek: Voyager*. USA: Paramount Pictures.
- Menosky, J. (Writer) & Vejar, M. (Director). (1999). Latent Image [Television Series Episode]. *Star Trek: Voyager*. USA: Paramount Pictures.
- Moore, R.D. (Writer) & Singer, A. (Director). (1993). Descent I [Television Series Episode]. *Star Trek: The Next Generation*. USA: Paramount Pictures.
- Moor, J. M. (2006). The nature, importance, and difficulty of machine ethics. *Intelligent Systems*, IEEE 21 (4), 18-21.
- Natsoulas, T. (1977). On perceptual aboutness. *Behaviorism* 5 (1), 75-97.
- Picard, R.W. & Picard, R. (1997). *Affective computing*. Vol. 252. Cambridge: Massachusetts Institute of Technology Press.
- Prince-Hughes, D. (2004). *Songs of the Gorilla Nation*. New York: Harmony.
- Rees, M. (2003). *Our Final Hour: A Scientist's Warning: How Terror, Error, and Environmental Disaster Threatens Humankind's Future in This Century-on Earth and Beyond*. New York: Basic Books.
- Russell, S. & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. New Jersey: Prentice Hall.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences* 3, 417-424.
- Snodgrass, M.M. (Writer), & Scheerer, R. (Director). (1989). The Measure of Man [Television Series Episode]. *Star Trek: The Next Generation*. USA: Paramount Pictures.
- Taylor, P. W. (2003). The ethics of respect for nature. *Environmentalism: Critical Concepts*, 1(3), 61.
- Wallach, W. & Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press, 2008.