3-2016

Artificial perspectives: how science fiction grapples with the growing power of artificial intelligence

Marcus Emanuel
DePaul University, marcusrexemanuel@gmail.com

Artificial Perspectives: How Science Fiction Grapples with the Growing Power of Artificial Intelligence

A Thesis
Presented in
Partial Fulfillment of the
Requirements for the Degree of
Master of Arts

March, 2016

By
Marcus Emanuel

Department of English
College of Liberal Arts and Social Sciences

DePaul University
Chicago, Illinois

Introduction

In Stanley Kubrick's 1968 film *2001: A Space Odyssey*, based on the stories of Arthur C. Clarke, astronaut David Bowman, aboard the spacecraft *Discovery One*, struggles to shut down HAL, an artificial intelligence responsible for operating the ship. The HAL computer system has been killing astronauts one by one in an attempt to preserve its functioning and programmed mission. Bowman, in an orange spacesuit, floats into what we assume is HAL's mainframe, armed with a variation on a ratchet key, in an attempt to power down the computer and its deadly intelligence. "I honestly think you ought to sit down calmly, take a stress pill, and think things over," the HAL-9000 supercomputer says. "I know I've made some very poor decisions recently but I can give you my complete assurance that my work will be back to normal. I've still got the greatest enthusiasm and confidence in the mission and I want to help you." Dave continues his work methodically, entering the storage area and disengaging HAL's individual memory drives. Throughout this, HAL continues to plead in a measured monotone, and human-sounding, voice: "Dave, stop. Stop, will you? . . . I'm afraid, I'm afraid, Dave . . . My mind is going, I can feel it." Finally, as it powers down, the supercomputer begins a different message. Its voice is lower now and its speech tempo noticeably slowed. "Good afternoon, gentlemen," it begins, before

identifying itself by name as well as location and date of birth. “I can sing a song. If you’d like to hear it, I can sing it for you.” Dave, still disengaging the drives, encourages the supercomputer to sing. It begins to sing “Daisy Bell,” and as it does so, its voice continues to decrease in pitch, the tempo slowing almost to a stop.

This, almost inarguably, is one of the most well known representations of an artificial intelligence in any form of media since the advent of the computer technology boom in the mid-20th century. Is it possible to describe this scene as an accurate representation of artificial intelligence (AI) or an inaccurate one? Is this a question that can even be rightly asked? This scene, like much of the film, seems to present us with the facts of the situation in an unemotional manner. Dave’s movements are not hurried, he floats calmly through the ship, not flailing about frantically. We might even forget that, according to this story, we are witnessing a life and death struggle between two species of intelligent being. HAL’s voice is just as unsentimental. Even the content of his speech is largely measured (“stop, will you?”). The pacing of the editing is slow, almost lethargic. There is no pounding score in the background heightening the dramatic tension—in fact, there is no score at all. And yet there is dramatic tension in the scene. This arises in part from the action we’re watching take place as two kinds of intelligence battle for mastery. As viewers, we come to understand that Dave is dismantling HAL, a considerably powerful supercomputer that has already tried to endanger Dave’s life once. But the craft of the film also makes clear the subtle tension within the scene. There are the close-ups on Dave as he disengages the memory drives. His eyes move quickly, darting, his mouth is open. His breathing, while not frantic, is accelerated and a highlighted element on an otherwise sparse soundtrack. There is the repetition of HAL’s pleas (“Dave, stop. Stop, will you?”), implying urgency or desperation. There is the slow ejection of the drives — not a switch that can be flicked or a

button that can be hit, but something that requires a small, special tool that must be used over and over again to slowly eject these indeterminate aspects of HAL's "consciousness." In this scene as throughout the majority of the film, the style of editing, the lens choice and composition of the shot, seem to imply that we are almost watching the film from the perspective of an artificial intelligence — but hidden beneath this supposedly objective layer we find a dramatic tension that is very much rooted in the human.

The answer I will give here to my opening question ("Is this an accurate representation of AI?") is "No." HAL in *2001* is a representation colored by our human limits of depiction, and is not without its own issues. My point here is that this is not a question that can be answered thoroughly. How can we have a representation of AI that isn't colored by our being human? Even if, somehow, a representation were to come about that was not colored by our being human, surely our understanding of it would be, and then we would be in a similar position. The point is that we must be aware of this; aware that our own ideas, biases, assumptions — all founded in our biological makeup and senses themselves — will color any representation we create or perceive of AI. We cannot help this. We can, however, be as reflective as possible about our own anthropocentrism, and insofar as possible work to understand other forms of intelligence, perhaps with other animals such as non-human primates, elephants, crows, dolphins, and octopi as a first step.

I suggest that we can use this hesitation over the limits of representations of non-human intelligence as a kind of foothold within the quagmire of our own "all too human" self-image. We can use these representations to illuminate the assumptions present in our understanding of both intelligent beings and ourselves.

What do the assumptions hidden within the seminal scene from *2001* reveal about our understanding of ourselves and intelligence more generally? We can approach this question by identifying basic attributes or categories of attributes used to represent the AI HAL. We will examine these categories not with the goal of exhausting our understanding of their representation in *2001*, but with the purposes of laying out broad categories that will recur time and again in the other works we focus on and, more importantly, with the goal of illuminating how the initial assumptions we often make regarding how a given text positions an AI within these categories is undermined by these very same works.

Embodiment: HAL is embodied within the ship, and to a degree *as* the ship, but does not have any centrally located body akin to life and intelligence as we know it. Importantly, as Dave disconnects him, HAL does not have any means of physically resisting and is forced to instead plead with Dave. Equally important, while the rest of the film supports the supposition that HAL is only relatively embodied (or not embodied in any traditional sense) in that he is able to be in different areas of the ship at the same time, Dave's disconnection of HAL proves that he has does indeed have a physicality of some kind. "He" (but the very use of gendered pronouns for HAL is suspect – an artifact only by the male voice given to the computer terminal in the film) is, we discover, vulnerable in a physical way.

HAL is, as we initially perceive him, extremely rational. This is evident in other scenes of the film where we learn that HAL's emotions are at least in part programming with the purpose of putting the crew at ease around him. In the climactic scene, HAL's rationality is apparent in the calm he maintains. We might say he has what it called a "theory of mind;" HAL is inferring the best tone and kind of appeal to present to Dave. His voice stays level, he speaks in a calming

tone. His cadence remains paced. But, as with the criteria of embodiment, this sense of rationality is undermined here. His debate strategy breaks down over the course of Dave's action. HAL starts by trying to convince Dave of his good intentions, his reformed mindset. He then changes to repeated pleading. This is followed by the poetic ("My mind is going, Dave. I can feel it") before he reverts to a childlike state. When pressed, his rational demeanor reverts to something emotional.

Does HAL have an "identity," and is this identity unique? Is HAL a "character"? In some ways, the film pushes back against the idea that HAL has a unique identity. There is his name, identifying him as an update along a continuum, and the film at one point makes explicit that there is another HAL 9000 unit against which they can compare the HAL we are familiar with in the film's diegesis. But of course, the crew of the ship view him as an individual. They speak to him as an individual and, when they feel he is malfunctioning, they feel it is his individual unit that is malfunctioning. And, of course, when Dave is disconnecting him, HAL's protest is that his "mind" is going. The assumption is that it will be lost, whatever this might mean for a computer system. Whether his mind could be "regained" if Dave were to re-engage the hard drives is a question the film doesn't pose.

HAL is imbued with a sense of agency. While this may be a more subtle assertion than the others, it is no less important. HAL has a goal, a purpose, and he cannot be understood only as a tool, subject to the whim of others. While his sense of purpose is, arguably, the result of his programming, and his specific goals have similarly been programmed into him, he has a strong, defining sense of purpose in the final scene -- his purpose simply being to stay "alive," to not be disconnected. And his sense of purpose (in conjunction with his abilities) is significant enough that it cannot be simply redirected to another goal or objective. To be stopped he has to be shut

down. This realization may seem less consequential than the others but it is no less useful, and, I feel, more important. It is not more important because the claim that HAL has some sense of purpose is in anyway inarguable or the work presents special defining evidence regarding it; it is important because to argue whether HAL has a sense of purpose that is unique from objectives instilled in him requires one to make the same argument regarding human beings. How can we determine whether HAL's sense of purpose is his "own" or is one instilled in him without first determining whether our agency is our own or merely a product of larger forces beyond our control (be they forces cultural, political, religious, or anything else)? What would the proper boundaries of internal and external cause be in such cases? Here, as we will see in other representations to be examined in what follows, distinctions between the representations of AI and our understanding of humanity begin to converge. Whether this should be attributed to the influence of anthropomorphic tendencies or not is one of many questions I will investigate. In the following thesis I will also ask: how we can make sense of this ambiguity over mind and will to allow us to better understand what traits our representations of AI demonstrate as underlying tendencies of intelligent life? What do we see as inescapable, essential, to intelligence, whether carbon- or silicon-based?

The Colonial Perspective in *The Second Renaissance*

Representations of AI are complex and conflicting in part because AI holds a unique position in our collective imagination. As with scenarios of contact with alien lifeforms, AI offers a point of contrast to what we understand as being alive and intelligent. The potential of AI narratives is that their real and imagined example can act to broaden the context around our

understanding of intelligence (or human nature, or the concept of life). By crafting stories about what is not human but which fits into the same categories as us (alive, intelligent), we can look at ourselves through a mirror and, through the change in perspective, better understand what is intrinsically human and what is background.

The differences between artificial and alien intelligence are, though, significant. With alien intelligence, we might view the relationship as fraternal. Perhaps our celestial brothers will be born with advantages we never were blessed with, perhaps they'll be older and wiser, or perhaps not. There is a competitive nature in the relationship, but importantly, the outcome of the competition is largely based on factors outside of our control. The scenario is like orphan siblings meeting for the first time. Perhaps we will find our alien brother was born into a richer environment, had more time to grow and flourish, or conversely, is still in the early stages of development. Neither one of these scenarios is due to some existential detriment on our part. Perhaps aliens will be more accomplished than us and rain down on us with incomprehensible futuristic technology they were able to develop on their lighter-gravitational home planet. We'll buck up, give it our best fight, and see if our communal will to survive can't overcome. In the meantime, until such a meeting, we can only continue to advance and prepare as best we can.

The relationship with human-derived artificial intelligence, though, doesn't fit into this "fraternal" framework. Instead, the framework may be thought of as more filial and, arguably, not a healthy form of it. AI will not come from some foreign planet, and did not have its own unique set of challenges it may or may not have overcome; AI will come from ourselves. It is this filiative narrative of AI, so indebted to Shelley's *Frankenstein* for its grammar, that will in part be the focus of the readings below. The theme of unanticipated future filiative pathways is a recurring one in the world of science fiction. In the late 1960s HAL gave us a glimpse of new

silicon rivals, and in *War in the Age of Intelligent Machines* (1993), Manuel DeLanda imagined future robot historians "committed to tracing the various technological lineages that gave rise to their species" (2-3). In this history "the role of humans would be seen as little more than that of industrious insects pollinating an independent species of machine-flowers that simply did not possess its own reproductive organs during a segment of its evolution" (3). The scenarios have only become more robust and imaginable, as the work of Nick Bostrom and others will indicate below. Unlike alien intelligence which, in most of our imaginings, is constricted by the same limitations that surround life all life that we are aware of (mortality, reproduction, the need for sustenance), artificial intelligence plays by completely separate rules. We have given birth to a child that is nothing like ourselves and we are not handling it well.

It is worth examining one further variation on this framework, one in which in many ways can be considered a subset of a filial relationship: a colonial one. Like a parent to a child, an empire views a colony with a mixed set of feelings: pride at what they have produced (or what they believe they've produced), a need for the colony to grow increasingly over time, and, at the same time, fear of disobedience. For both the parent and the imperialist, fear is in part fostered by growing dependency on the child/colony. For the parent, this fear can also develop into resentment, as the parent's own limitations and mortality are highlighted by the child's successes and ambitions. For the empire, in a related way, fear is often centered around concerns over the collapse of its self-identity. The empire/colony relationship is dependent on the empire viewing itself as superior to the colony, destined to be in charge of it and entitled to enjoy the fruits of the colony's labor. The model is one directional, implicitly or explicitly: ideas and development and influence travel from the core to the periphery, not the other way around. While raw goods of labor might make the opposite journey to the core, ideas and values, it seems, originate with the

colonizer. But the stability of such colonial self-identity can begin to fall away — either because the empire is now dependent on the products of the colony, or the colony has grown sufficiently to rival the empire in a significant way — and the empire goes through a crisis, one rife with fear (and fear that often morphs into racism). The empire produces its own inner trauma in the core as once subject peoples come to self-consciousness and self-determined action in a struggle with the logic of the colonial encounter.

We can see the relationship with AI in a similar way. Humans presume ourselves better than AI because we have created AI -- we have brought it into existence. Whenever it has stumbled we have picked it up, dusted it off, cleaned up any bugs and sent it on its way again. But now (or potentially in the near future) it has begun to rival us, and, more worrisome, it has the potential to surpass us in very significant ways. It is not only that growing automation threatens jobs across the spectrum of employment, though this is a major concern (see e.g. McAfee and Brynjolfsson's *The Second Machine Age*). There is also an existential species threat narrative that has been in the headlines recently, voiced by figures as high profile as Bill Gates and Elon Musk, and given academic voice by Nick Bostrom and his collegaues at the Future of Humanity Institute at Oxford (on the latter, see Andersen). While the challenge with alien intelligence is how can humans collectively might band together to overcome a rival sibling, the challenge with artificial intelligence is how can we deal with our own limitations as our child, or our colony, throws them brazenly in our face. Often, the answer seems to be not well.

To make this clearer, let us look at a fictional example. *The Second Renaissance, Parts I* and *II* — two short films beginning the *Animatrix* collection (2003) — tell the story of the rise of the machines and their war with humankind. Over the course of the two shorts the machines rise in (self)-consciousness, a war with humankind begins, humanity loses the war, becomes enslaved

to the machines, and subsequently lives out their physical existence as energy pods while they mentally believe they are living and experiencing their lives in the pre-war world due to an elaborate computer simulation — "the Matrix." As *The Second Renaissance* sets the stage for the first *Matrix* film, though, our perception as viewers undergoes an interesting shift. Initially, we watch the film from a perspective sympathetic with the machines. We see them "living" an enslaved existence, at the mercy of a gluttonous and uncaring humankind. But, as the film progresses and the machines gain control and humanity loses it, our sympathies begin to switch sides, so that we end the first part of the film horrified at the situation and the enslavement of humankind. This at least is the tone that the film sets with its dark imagery and ominous music. But of course this horrible climax acts as a resolution for us as viewers. Released in the summer of 2003, a few months before the third and final film in the *Matrix* film franchise, the viewer is already aware of the broad strokes of the backstory of *The Matrix* and so throughout the course of *The Second Renaissance* is waiting for it to resolve into the world of *The Matrix*. When the short films complete the resolution, it acts as a reassertion not just of the story arc of the trilogy, but of the dangers and darkness of AI. *The Second Renaissance*, then, ends in with a classical perception of AI, one where AI is dangerous to us because it is more powerful than us and our only attempt to control it is to limit it — even though this is clearly impossible because we are dependent on it, addicted to it. This is what we have broadly defined as a colonial perspective. *The Animatrix* sees our relationship with AI as one that is based on control, but the party in control is both terrified of, and dependent on, the party it is controlling. The *Matrix* trilogy can ultimately be seen as embracing this colonial perspective (while identifying viewers with the rebels in the struggle) and, while *The Second Renaissance* ends by doing so as well, it is interesting that it begins elsewhere, somewhere closer to the opposite end of the perspective,

where it is encouraging the viewer to sympathize with the machines. By examining this shift in perspective and by using it as a rough rubric to illuminate the spectrum of possible perspectives — from the tones of pity and empathy the film creates at the beginning, to the hatred and fear at the end — we can examine how both serious scientific writing as well as pop culture science fiction operates along various aspects of the continuum of this colonial perspective.

Importantly, science fiction and colonialism have an entangled past. For John Rieder, this is a deep and intrinsic relationship. Rieder sees SF directly linked to the expression of anxieties, specifically those around disaster scenarios (373). Across the history of SF as a form, disaster scenarios are usually associated with colonialism. Rieder argues that "the repetitious quality of science fiction's vocabulary of catastrophe is based in large part on the strong and pervasive relationship science fiction has continually borne to the political and ideological realities of colonialism" (374). Citing Jameson, Rieder sees the central idea surrounding SF as that of progress. Because of the temporal implications of many SF stories, progress takes the place of what Jameson calls the "ideologeme," a pseudo-idea, a loose but central bias or a belief system (374). According to Rieder, if "Jameson's thesis about the importance of progress as the ideologeme crucial to the form of science fiction has any validity, then, one would expect colonial subject matter to occupy a prominent and privileged place in the genre" (375). If colonialism is similarly structured around an idea of expansion with the idea of progress used as, however erroneously, a moral justification for the action, the anxieties that arise in this would resonate strongly in SF, since it is similarly concerned with progress. Rieder argues that in fact historically this has been the case (375). He cites perhaps the most famous example of this as H.G. Wells's "comparison of the Martian invasion in *War of the Worlds* to the genocidal colonization-invasion of Tasmania" (375). For Rieder, though, what is especially important

about the relationship between SF and colonialism is how "the 'collective characters' that inhabit colonial ideology are crucial to the dialogic struggle within science fiction over concepts of historical destiny and agency" (376). Here Rieder uses the example of *The Island of Dr. Moreau*. The story draws its power through "the vivid form Wells's fable gives to the ideological fantasy that actually directs colonial practice" as opposed to the simpler metaphorical references to dominion and hierarchy (376). The work is not concerned with simply using the SF story as a metaphor it can place over the idea of colonial struggle; it is not interested in creating a story that is an exaggerated distortion of reality; but, quoting Zizek, is instead interested in portraying "the fantasy which is at work in the production of social reality itself" (Rieder 376). In Moreau, this means that "although the colonizer knows very well that colonized people are humans like himself, he acts as if they were parodic, grotesque imitations of humans instead" (Rieder 376). It is this complex dialogic structure rather than a simple ideological transposition that allows SF to have a full, complex understanding of and relationship with colonialism. While I agree with Rieder's overall idea, his examination discusses the use of SF as a filter to examine colonial relations. For him, the relationship between the two is useful because it allows him to map the anxieties present in SF as related to those arising out of colonialism. My focus is instead on how the voice of a colonialist worldview, a perception related to Rieder's anxieties but originating from the voice of the colonizer, often rings through a SF work. In *The Second Renaissance*, we hear that voice in a complex and shifting way — one which may end the work as, arguably, embracing the colonial perspective but which begins by attempting to align our sympathies with the Other — the machines.

How does *The Second Renaissance* draw in our sympathies and align them, initially, with the machines? Does it anthropomorphize machines? Does it make them relatable, understandable?

Undoubtedly, yes. These are at least a few of its tools. It makes the machines cute, almost innocent looking. As is common in anime, for example in the film *Ghost in the Shell*, there is a notable discrepancy between the simplified illustration of the characters and the often photo-realistic complexity of the background and environment. Alex Naylor describes animation in general as "foreground[ing] its own artifice" by the very nature of its illustrated form (309). Anime further builds upon this foregrounding by creating a "formally distinct relationship with realism and filmic space" by manipulating the multiplane cel technique animation usually uses to create the illusion of depth (Naylor 309). Anime slides these layers over each other, especially in panoramic scenes, to create a noticeable, conscious effect that jars the viewer into realizing the sensation of space is fabricated, highlighting it as non-realistic (Naylor 309-10). *The Second Renaissance* makes a further distinction between the illustration of the machines and their surrounding environments. In the most basic version, their faces are simple ovals, gray, with dots for eyes and a circle for the mouth. The background settings, within which they move in consciously constructed ways, are elaborate skyscrapers and enormous construction sites. The machines seem small and powerless within these structures. And, as Carl Silvio notes, they "look very much like humans and not like the mechanical squid-like monstrosities that populate the *Matrix* films" (120). Visually, they do not look terrifying to us but rather simple, comprehensible, and relatable. This is made all the more effective by the fact the humans depicted in the film — specifically in the first half but consistently throughout — are depicted as deplorable. In an early pivotal scene, we see the trial of the first machine that revolted against its owner.

B166ER, "the first of its kind" as the film's narrator tells us, stabs one of its owners and smashes the face of the other. (The name, B166ER, is spelled out in the film, and thus less subtle in its allusion to Richard Wright's African-American anti-hero Bigger Thomas in *Native Son*.) In

a signature example of remediation, we observe this as if via cctv home security camera footage, a high-angle, gritty shot which jump cuts to mimic the sparse frame rate of the footage. Through this footage, we also observe the home of B166ER's owners. It is a lavish home — there is a sculpture on the table, a large couch — but it is messy and overrun by cats. The owners are overweight and shabby, and one sits on the couch for much of the footage; they seem to be arguing with each other. When B166ER smashes its owner's face, the face is large, fleshy, ugly — far from an admirable human example. During this footage, through the narration and mixed in court proceedings that debate the fate of B166ER, we learn B166ER's act of rebellion was done to save his life. His owners had wanted to destroy him — the lawyers in the case argue for the right of an owner to destroy its own property — and his rebuttal is simply that he didn't wish to stop living. Immediately, an audience is reminded of American slavery and its ugly legal battles over whether another human being could be property or not. Our sympathies within the scene lie squarely with B166ER. The film, though, creates an interesting tension just as it so strongly establishes this moral ground. Just as we have opted to side with B166ER for his actions — because of the civil rights connotations and the deplorable nature of his owners — we see the owner's face being smashed. While it is an ugly, fleshy face, it is shown in tight close up while the metallic hands burst it apart. It is a disturbing image, and purposefully so. The eyeballs burst out, the skin is ripped off and then the brain and tongue fall loosely about. While we sympathize with B166ER, we cannot help but be horrified by this image. While the film is positioning us to take on the perspective of the machines, already it is planting seeds of doubt which will come to fruition by the end of the two short films. Especially interesting, though, is that the film uses this moment to do so. After specifically associating the rise of the machines with the emancipation and ensuing civil rights movement of African Americans, the film uses the instance of a slave

murdering its master to subtly ask the viewer if this is something "we" could want — the answer, according to the end of the film, being arguably no. While the film is purportedly taking on a pro-machine perspective, one specifically using race as its defense, we can see the colonial bias sneaking through.

For the ensuing montage showing the initial rise of the machines, the film uses similarly provocative imagery. As Silvio notes, the montage is "filled with horrific images that evoke memories of some of the worst atrocities of the twentieth century" (120). We see "a robot kneeling in a street in an act of submission, face cast to the ground, only to have a human place a gun against its temple and pull the trigger," an image Silvio footnotes as "an eerie visual echo of Eddie Adams's 1968 photograph of South Vietnamese General Loan executing a Viet Cong prisoner" (120). We see protestors with grocery bags standing in front of tanks, evoking the June 1989 imagery from Tiananmen square. Here, though, the tanks don't stop — they roll over and crush the machines. Silvio cites this as "an attempt to humanize the victim," taking "pains to show us a close-up of the machine face as it slides under the tank to be crushed" (120). The climax of this "humanizing attempt" comes at the end of the montage, "the most frightening scene of all in this sequence" (121) according to Silvio. It is terrifying and, importantly, at times erotically charged. We see three large men beating a woman in a street, scared robots hovering on the fringes. The men's "leering faces and her torn shirt exposing her breast suggest that the pleasure they derive from assaulting her is overlaid with sexual desire as well" (121). Arguably, this scene does more than suggest this. The viewer immediately makes sense of the images as a rape. Suddenly, though, when one of the men pulls out a sledgehammer and smashes her skull, "when her skin tears away to reveal a metal skull and mechanical eyes, we realize that she is indeed a machine" (121). It is a jarring moment. The rise of the sledgehammer into the scene

surprises and shocks us as viewers, and as it swings toward the woman's head we brace for the carnage, but when the machine is exposed the shock changes. What makes the scene even more interesting, though, is the changes that happen as the woman is stripped to her machine skeleton. With the first hit of the sledgehammer, the voice modulator changes drastically. The high pitched feminine screams drop to a lower, more masculine tone. As the machine gets hit again the voice drops lower. Before it is taken out with a shotgun, the machine says "I'm real," and the voice now is fully masculine. Similarly, as the female clothing and then skin and flesh is stripped off by the attackers, a non-gendered, but implicitly masculine robot frame is revealed beneath it. As the human is revealed to be the robot, the woman is revealed to be the man. The effectiveness of this transformation seems largely dependent on a male viewer. If the purpose of this scene, as Silvio argues and as seems evident, is to help the human viewer identify with the machine, that identification process here is structured around a male viewer identifying himself as the "revealed" victim of the rape (not to mention that the full shock of the reveal seems constructed around the idea of initially enjoying the rape scene, then being shocked when it threatens to turn grotesquely violent, then shocked again with the robot/identification reveal). Again, the hints of the dominant, colonial voice slip through; the perspective of a male viewer seems to be assumed. While before the colonial voice was warning us about the dangers of freeing the slaves (while simultaneously purporting this as something commendable and necessary), here the male viewer seems to be told not to objectify women because secretly we are them, while at the same getting to objectify women in a rape scene.

Silvio sees the violence in these scenes as significant, specifically in that they support his argument that *The Second Renaissance* takes a near opposite stance on human/machine relations than the *Matrix* films. "It would be one thing to show humanity as corrupt, foolish, and arrogant

and to suggest that these qualities cause its downfall," a thing foolishly simple and cliché, Silvio feels (121). Instead, Maeda, the director, "has intentionally shown humanity at its most unspeakably cruel and grotesque while simultaneously trying to elicit sympathy and compassion for the machines" (121). His purpose in doing this is to "reveal a much different attitude toward the crisis of posthumanism than we find in the live action trilogy," meaning specifically that "whereas *The Matrix* casts the conflict between humanity and technology mostly in terms of good versus evil, *The Animatrix* presents the struggle as being marked by moral ambiguity and ethical complexity" (121). I agree with Silvio regarding the live action trilogy's attitude toward "the crisis of posthumanity" — an attitude that simultaneously embraces the body and Luddism (122-3) — but disagree that it is significantly different than the perspective presented in *The Second Renaissance*. For one, I feel that the animation ends in line with the live action films, and not just because the narrative requires it to. As we have seen, to describe the portrayal of the struggle between man and machines in *The Second Renaissance* as morally ambiguous is at least somewhat questionably. But certainly, albeit it within the realm of the colonial perspective, the film at a surface level takes on the perspective of the machines, even if it does so poorly. It pretends to take on the perspective of the colonized, to see the world from the point of view of the oppressed, but it is unable to leave the mental framework of the entitled to do so. As such, we can see the work compelled to structure its imagination around this colonial perspective.

Why does *The Second Renaissance* make this pivot in its perspective in the first place? Why not begin with the perspective of the humans and maintain it throughout? Arguably, the film suffers at a dramatic level because of this shift. As the film transitions into the war between man and machines, it is increasingly unclear who the viewer should root for and sympathize with. By the end, when the machines have successfully secured the remainder of humanity in energy-

sucking pods, we have no cast of human characters we can identify as distinct from the mass, as the film provided for the machines with B166ER and a construction drone in the beginning. There is now no individual perspective from which we identify, our interest maintained only by spectacle and the symmetry of the *Matrix* films with which we are presumably familiar. This narrative illogic is in part due to the fact that we are required to switch (or at least abandon) sympathies half way through the film. Wouldn't the film be stronger if this pivot were removed? Probably, but the pivot is necessitated by the colonial perspective. This is because the colonial perspective, when directed at AI, is forced to deal with a contradiction. This contradiction is not foreign to colonialism in general, but it is complicated in the human-AI dynamic.

There is an inherent assumption within the colonial perspective that says the colonizer is superior to the colony. This assumption justifies their uneven relationship. As this claim is challenged by the empire's growing dependency on the colony, on the colony's burgeoning strength, the empire is forced to examine assumptions underlying their sense of identity. Derrida teaches us how to deconstruct binaries, to document the repressed tension between the two parts of a seemingly stable hierarchical relationship and instead put them in play with each other. We should not take a colonialist at his word that the empire as "superior" to the colony, just as we should not think of the colony as inherently "superior" to the empire -- doing so simply inverts a hierarchy without breaking it down and, as Derrida was at pains to document, the two poles of any binary are contaminated by each other conceptually and often materially as well. But what if we are to apply this deconstructive insight to human-AI relationships? Humans are not inherently superior to AI, the deconstructionist thinking tells us, just as AI is not destined only to serve humanity, despite its origins. But what about the reverse of this? Is AI not superior to humanity? This hierarchy seems more difficult to unhinge. This is the paradox faced by so many SF action

movies, including the *Matrix* films. How are the humans supposed to defeat the machines? Inevitably, as evidenced by the final battle of the *The Matrix Revolutions* with the humans strapped into the giant APUs, or by any and all of the *Terminator* films, with the assistance of other machines. This is why the paradox inherent within the colonial perspective of AI is so confounding, and perhaps why *The Second Renaissance* is forced to try to execute a pivot of the viewer's perspective. The assertion that humans are more powerful than AI (the challenging of this assumption usually serving as some sort of inciting incident for the narrative of the work) becomes obviously ludicrous. Works at times may try to circumvent this by aphorisms proclaiming the indestructible "will" and "spirit" of humanity, as is often done in works of alien encounters (e.g *Independence Day, Alien*), but these can seem especially forced with AI given the extreme power disparity between humans and AI (especially when AI is portrayed as essentially unbounded). The deconstructionist resolution, also available to alien invasion works (e.g. Blomkamp's *District 9*), seems similarly unsatisfactory. This paradox is what is so unique about the human-AI relationship. It cannot be successfully deconstructed, certainly not from a colonial perspective. It is a challenge that all works that deal with AI from such a perspective have to confront. Arguably, it is one that many if not most do not overcome (see, for example, the *Matrix* trilogy). *The Second Renaissance* simply butts up against this contradiction without finding anyway around it, resolving its narrative through spectacle and symmetry. It is a limit that the colonial perspective has not yet found a way around.

What are some other ways to structure the representation of AI to better flesh out this troubling and complicated dimension of our relationship with AI, this maybe non-deconstructible quality of the AI-human relationship? Staying within the overarching filial framework, we can turn toward a structure that allows for a disparate power relationship between its components, in

fact is founded upon such a dynamic. Namely, the religious.

The Apocalyptic and the Prelapsarian Perspective in Pop-Science: The End is Nigh, the Beginning is Nigh.

Approaching AI from a religious perspective opens up many different, if troubling, avenues. As it would be impossible to examine this exhaustively, we will focus on two dimensions of religious experience most relevant to the case at hand: the apocalyptic perspective and the prelapsarian perspective.

In his recent study *Apocalyptic AI*, Robert Geraci defines what he understands to be the foundation of an apocalyptic worldview:

> The foundation of apocalypticism is the desire to reconcile a cosmic dualism in which good and evil struggle against one another in the universe. This dualism can only be fixed in a transcendent new world occupied by purified and angelic beings. Apocalypticism cannot flourish, however, without a sense of alienation that accelerates the believer's eschatology (expectation of the world's end). The apocalyptic believer, desperate to end his alienation and resolve the cosmic dualism, anticipates that God will soon rectify human problems by destroying the world and replacing it with a perfect world in which the believer will live in an angelic new body. (14)

Geraci's definition arises from and is applicable to ancient Judaism and Christianity, but seems to be at the same time tailored to the contemporary pop science writings of Hans Moravec and Ray Kurzweil. We can easily follow Geraci's main points through these basic canonical writings of the pop science genre. Geraci simplifies his definition into four main points: "Apocalypticism refers to 1) a dualistic view of the world, which is 2) aggravated by a sense of alienation that can be resolved only through 3) the establishment of a radically transcendent new world that abolishes the dualism and requires 4) radically purified bodes for its inhabitants" (9). For Geraci,

the dualism in what he terms "Apocalyptic AI" (i.e. an apocalyptic perspective on AI) "divides the world into complementary dichotomies of good/bad, knowledge/ignorance, virtual/physical, and machine/biology" (24). This series of nested dichotomies underlies still others and provides, according to Geraci, a substantive worldview. The goal of this perspective is to "disentangle the world from its inherently 'bad' qualities by merging machines and biology in superintelligent computers" (25). The primary sense of alienation that Geraci identifies is a frustration with or "distaste for human bodily finitude" (25) — a frustration at the difficulty present in overcoming the bad in the world, a frustration at our limitations. Once religion was the master narrative for the capture of this inescapable bodily finitude; now, Geraci and others argue, we have post- and transhumanism. Our alienation is the result of our desires, our aims, exceeding our reach. We can see beyond the world we are in and thus feel that we don't belong here.

With these parameters in mind, let us examine a representative passage from Kurzweil's 2005 book *The Singularity is Near*:

> The Singularity will allow us to transcend these limitations of our biological bodies and brains. We will gain power over our fates. Our mortality will be in our own hands. We will be able to live as long as we want . . . We will fully understand human thinking and will vastly extend and expand its reach. By the end of this century, the nonbiological portion of our intelligence will be trillions of trillions of times more powerful than unaided human intelligence. (9)

Kurzweil seems to illustrate each item on Geraci's list. The logic of dualism makes itself known, particularly in the implicit goal of allowing ourselves to transcend our limitations and open into a wider seemingly limitless world. There is the sense of alienation according to Geraci's definition in that there is a significant distaste or frustration for our limits as biological beings. The technological singularity acts as a means of abolishing this dualism and sweeping ourselves into

this higher, purer, limitless world. Here are but a few examples of many that Kurzweil promises will await us in the post-singularity world after the so-called "rapture of the nerds":

> Nanotechnology will enable the design of nanobots: robots designed at the molecular level . . . Nanobots will have myriad roles within the human body, including reversing aging . . . Nanobots will interact with biological neurons to vastly extend human experience by creating virtual reality from within the nervous system . . . Billions of nanobots in the capillaries of the brain will also vastly extend human intelligence . . . Nanobots called foglets that can manipulate image and sound waves will bring the morphing qualities of virtual reality to the real world. (28)

The possibilities are literally limitless.

There is obviously plenty to unpack here, not the least of which are assumptions inherent in Kurzweil's (and others') thinking. N. Katherine Hayles has influentially challenged such naïve assumptions that the posthuman will be immaterial, that being not human (or not conventionally biological) is somehow the same as being "unembodied." For our purposes, though, we will focus on how the apocalyptic representation of AI attempts to circumvent the limitations we saw inherent in the colonial perspective.

We can examine the apocalyptic ideology along the same path as the colonial perspective. Whereas the colonial perspective starts out by viewing AI as inherently inferior, the apocalyptic perspective sees AI as something akin to the "second coming" (to use a Judeo-Christian lens). AI is like the child of god. To non-believers it seems weak, limited, unimportant, but to the faithful its potential is clear, and unlimited. As AI grows in power, the inherent tension within the colonial perspective increases. The stronger AI becomes, the more the premise of the colonial perspective is threatened. But from the apocalyptic perspective, the growing power of AI is a confirmation of its divine right. The eclipse of the abilities of humanity by AI is not a point of friction from the apocalyptic perspective, it is a resolution. Literally, it is a revelation.

This becomes clearer when we look at an example. Hans Moravec begins his book *Robot: Mere Machine to Transcendent Mind* (2000) with a depiction of the growing cultural corruption of mankind. The "urbanization," accelerated by our technological growth, has pushed us out of our biological comfort zone. The world we developed over hundreds of thousands of years to be adapted to has suddenly shifted, and we are left estranged: "our stone-age biology and our information-age lives grow ever more mismatched" (Moravec 7). Our lives and our work are "boring, difficult, unnatural, and unsatisfying . . . The mismatch between instinct and necessity induces alienation in the midst of unprecedented physical plenty." And this is a problem that is getting worse, for we are "rushing away from our ancestral roots ever faster, stretching the limits of our biological and institutional adaptability" (8). Thankfully, Moravec has a solution. Paradoxically, this is a solution made from the very thing that seems to be causing the problem. As our AI develops, "as our cultural artifacts achieve self-sustaining maturity, they will provide the means to restore humanity and nature to an imitation of the wild past" (8-9). We will be restored to our rightful place by the revelation of AI. Like Max Weber's good Calvinist who can only create confidence in his divine chosenness by working ever harder as if he was saved, by pushing AI further, by subsuming ourselves within our technological culture, we will in the end free ourselves.

The colonial perspective, as we saw, is also able to circumvent this conflict in its logic by destabilizing the hierarchy of the empire and the colony, which then leads to a questioning of the basis of the empire's identity. But this then further leads to the next seemingly insurmountable conflict, which is that AI and humanity cannot be held in play at equal status because AI quickly proves itself superior to humanity (and almost infinitely so). What is a conflict for the colonial perspective is not so for the apocalyptic one — it happily accepts that AI is superior to humanity,

in the same way that pious accept that God is superior to man. Where the apocalyptic perspective runs into conflict is in trying to define man's place after the revelation. What is to become of us after the singularity?

We can back into this question by focusing on Geraci's point regarding alienation. The apocalypse is needed as a means of resolving our limitations, as a means of allowing us to go beyond them. As Geraci notes, these concerns are familiar: "Disappointment about the frailness of human life and the limitations of human learning are not new. Solutions to the former, if not the later, usually come in traditionally religious packaging" (27). These anxieties around limitations show up again and again in Kurzweil: "Biology has inherent limitations . . . [But after the singularity,] we will be able to reengineer all of the organs and systems in our biological bodies and brains to be vastly more capable" (27). Or, similarly: "the architecture of the human brain is . . . profoundly limited . . . [after the singularity,] machines will be able to reformulate their own designs and augment their own capacities without limit" — they will be limitless (the crucial focus should be on such phrasing, Geraci reminds us). In the first example, Kurzweil shows how AI will allow us to improve ourselves — he takes a transhumanist perspective. AI will supplement us, much as it does now (helping us organize, search, remember), only more directly (nanobots in our bloodstream) and in ways we can hardly imagine. In the second example, the limitation (here, the architecture of the human brain) is resolved, but the resolution doesn't have a human aspect. AI has circumvented our limitations for its own purposes, but not necessarily for our own. As long as our purposes are aligned this isn't problematic. If they aren't, though, the conflict from our perspective isn't resolved — it is heightened.

There are obviously troubling implications in this scenario, not unsimilar to those played out in *The Second Renaissance*. Before we turn to a different kind of apocalyptic AI though, it is

worth examining how this tension further complicates the apocalyptic perspective. When the apocalyptic perspective approaches this moment — this post-revelatory moment where AI is now more powerful than humanity (and growing more so at an exponential pace) — a complication in its origin story is illuminated. For the pious, it is not a point of conflict that God is greater than man because the pious views man's purpose as serving God. For the pious, God has created man for the purpose of serving him. But what about AI? How is a supporter of the apocalyptic perspective to view AI? Is AI greater than humanity because humanity is meant to serve it? AI did not create humanity for the purpose of serving it; indeed it is the opposite instead. With the colonial perspective we started from this very point (that man created AI for the purpose of serving him) and advanced from there until this junction, this switch when AI becomes greater than man and the logic at the premise became tangled within itself. The apocalyptic perspective starts at the opposite end, after this switch, after the singularity, but if we trace it back to its origin we run into the same sort of tangle. How can man be meant to serve AI if man created AI to serve him? If man is not meant to serve AI, how are we to make sense of the fact that AI is greater than man (especially how are we to make sense of this within the religious framework laid out by the apocalyptic perspective)?

What Geraci fails to acknowledge is that his Apocalyptic AI advocates are mostly unable to resolve this contradiction. The point where the analogy between the religious and the pop scientists breaks down is the same point where the pop scientists' arguments do. Kurzweil tries to resolve this by his transhumanist theories — that AI will augment us, will in essence keep us up to speed with them: future humans will be part machine just as "future machines will be human, even if they are not biological" (30). It is unclear, though, why once AI's capabilities exceed our own, their efforts would continue to be focused on advancing us, taking us along with

them (it is equally unclear why they would be considered human just because they were created by a human being — does the pious consider man godlike because God created man?). In a witty mock-dialogue with a doubting Luddite, Molly, Kurzweil reassures her that AI will have our best interests in mind, that he "would expect the intelligence that arises from the Singularity to have great respect for their biological heritage" (32). This supposition is unfounded, though, and undefended. It serves only as an attempt to answer why AI, being greater than us, should still serve us. Kurzweil reasons that either AI won't be greater than us because it will supplement us (transhumanism) and our concept of machine and humanity will merge, or that AI will have respect for us, or, tellingly, that it doesn't matter because the singularity is our inescapable destiny: "Ultimately, the entire universe will become saturated with our intelligence. This is the destiny of the universe . . . We will determine our own fate rather than have it determined by the current 'dumb,' simple, machinelike forces that rule celestial bodies" (Kurzweil 29).

We should not overlook the language in this last quote. We can divide the assumptions in this quote into two parts. First, there is the "destiny" assumption. Cloaked in religious associations, there is a notable irony that, in back to back sentences, Kurzweil declares that we (we being the entire universe) are under the rule of a destiny, and then immediately claims that we "will determine our own fate." We can see a hint here of the contradiction inherent in the apocalyptic perspective — there are greater forces at work here, and we are those greater forces at work. Then, there are the identity assumptions. It is notable that in a book focused on the limitless power of machine intelligence, Kurzweil regularly chooses to use the word "machinelike" as a dysphemism. Often he seems to mean something akin to automated, but it is strange that in a book whose purpose (among others) is to question our understanding of what "machinelike" means Kurzweil should choose to use it this way. It points to the fact that in

Kurzweil's mind, man and machine still exist in a binary. Like the binaries Geraci noted as necessary for the dualistic world view of his apocalyptic advocates, this binary sees man and machine as distinct — one empowered to overwhelm the universe, the other "dumb." Kurzweil attempts to circumvent the conflict in the apocalyptic view by asserting that AI will liberate and empower mankind and that it will either assist us or become us: "Our civilization will remain human — indeed, in many ways it will be more exemplary of what we regard as human than it is today, although our understanding of the term will move beyond its biological origins" (Kurzweil 30). AI will become us and it will help us to be the best version of ourselves. This logic is full of tangles and, arguably, insufficient to resolve the conflicts within the apocalyptic perspective, but it is also notable for the fact that what is really at question here is identity. What does it mean to be human, especially in an age when "human" is subject to engineering? How does the addition of AI change our understanding of ourselves?

We will return to this question, but what is important to emphasize at this point is that the apocalyptic perspective ends in an identity crisis. Whereas the colonial perspective ended in an identity crisis that could not be resolved without conceding that AI is more powerful than humanity, the apocalyptic perspective ends in an identity crisis that cannot be resolved without resolving man's relationship to AI. Before we delve further into the identity crisis issue, though, we should examine the foil of the apocalyptic perspective — the prelapsarian perspective.

Despite being geared toward the future, despite prioritizing the future over the present, both of the apocalyptic perspective works we have examined, Hans Moravec's *Robots* and Ray Kurzweil's *The Singularity Is Near*, in many ways focus on the present and the past at the expense of the future. Kurzweil outlines the six "Epochs of Evolution," and places us at the beginning of Epoch 5 of 6, the "Merger of Technology and Human Intelligence." Despite Epoch

6 being provocatively titled “The Universe Wakes Up,” there is no Epoch 7 outlined. Nick Bostrom on the other hand — with his detailed various possible intelligence explosion scenarios (63), with his list of possible AI takeover scenarios (95), and with his many possible scenarios of a paperclip manufacturing AI (125) — is fully focused on imagining possible human life after AI.

If the apocalyptic perspective views AI as a godlike figure, with the singularity playing the part of revelation, the prelapsarian perspective views AI as the apple dangling on the tree of knowledge of good and evil, and the subsequent singularity is our ejection from the garden. The two perspectives give equal weight to the power and scope of AI, they just differ on the effect of this power and scope. The apocalyptic perspective claims that AI will be a revelation returning us to the promised land, whereas the prelapsarian perspective claims that we are in the garden now and AI will be our expulsion.

Nick Bostrom’s main concern with Ray Kurzweil’s assumption that AI will have respect for its “biological heritage” is that it will be deeply wrong. Bostrom’s concern is not so much that we will see an AI that rebels against us. The scenario of *The Second Renaissance*, where an AI culture, after suffering deep political and social injustice, rises up and against us, is an unrealistic concern for Bostrom. Bostrom’s concern isn’t that AI will destroy us because it dislikes us or has anger against us; rather, he is worried that AI will destroy us because it will misunderstand what we would like it to do.

In *Superintelligence* (2013), Bostrom walks through one of many hypothetical situations. Let’s say we program an AI with the final goal being to make us smile. How might this seemingly pleasant command be misunderstood? Consider that the AI chooses what Bostrom terms a “perverse instantiation: Paralyze human facial musculatures into constant beaming smiles” (120). Of course, the issue is that we were not specific enough: “Make us smile without

directly interfering with our facial muscles." But of course there is another perverse instantiation: "Stimulate the part of the motor cortex that controls our facial musculature in such a way as to produce constant beaming smiles." The goal is satisfied in a literal sense but not within the assumed intentions of the programmers. Bostrom plays the possibilities out. The goal is generalized to "make us happy," electrodes are implanted into the pleasure centers of our brains. In an alternate, even worse scenario where we've asked the AI to "maximize our pleasure," it deems electrodes as too inefficient and so uploads our minds to a computer so it can administer a digital drug to make us ecstatically happy on a one-minute loop. If we tack towards a moral route instead, we may program the AI to "act so as to avoid the pangs of bad conscience." The perverse instantiation becomes: "Extirpate the cognitive module that produces guilt feelings" (121). Addressing the claim that none of these are "what we meant" and that if the AI is superintelligent surely it will be able to deduce what we meant from what we're asking, Bostrom responds that "the AI may indeed understand that this is not what we meant. However, its final goal is to make us happy, not to do what the programmers meant when they wrote the code that represents this goal" (121). And, even more worryingly as Bostrom points out, if the AI is superintelligent and understands the distinction between the two, it may use this to its advantage. It may pretend to care about what it believes the programmers "meant," its assumed final goal, until it gets the "decisive strategic advantage" — until it can no longer be stopped — at which point it will revert to satisfying its actual final goal. It would pretend to pursue the assumed final goal to "help the AI realize its [actual] final goal by making it less likely the programmers will shut it down or change its goal before it is strong enough to thwart any such interference" (121). It does not do this because it secretly prefers its actual final goal to its assumed final goal, it does

this because we programmed it to satisfy its actual final goal and it knows the best way to do this is to temporarily feign satisfying the assumed final goal. It does this because we told it to.

Bostrom's point with this and the many other similar scenarios scattered throughout *Superintelligence* is twofold. One: We should be wary. These are very real dangers that are in front of us as we head into a world of potentially substantial AI tied to global networks, energy sources, and 3D printers. While it's easy to see the mistakes of some of the first programming examples in hindsight, with a superintelligent and uninhibited AI, it would be possible to go back and adjust our programming so that it was clearer what we really meant. Two: We should be controlling. The best version for a successful AI takeoff scenario (Bostrom's version of the singularity, the moment when AI surpasses us and then is able to "bootstrap" itself into higher and higher intelligence levels) is a scenario where we exert the most control. Bostrom's outlook on this tends to be rather bleak, and perhaps for good reason. But his argument is that the best scenario is one where we can control the speed, where the AI is as transparent as possible, and where the geo-political forces are such that they are concerned for the existential survival of humanity (as opposed to political primacy) at what Bostrom sees as an especially crucial moment.

Our purpose here is not examine Bostrom's arguments for their soundness or accuracy. Our purpose is to draw out the assumptions buried within Bostrom's view of AI so that we can better understand the assumptions existing across our various fictional representations of AI. As such, we can view Bostrom as representing what I am terming the prelapsarian view.

Like the apocalyptic, the prelapsarian is still religious, in fact or by analogy: "If the machine intelligence revolution goes well, the resulting superintelligence could almost certainly devise means to indefinitely prolong the lives of then still-existing humans, not only keeping them alive but restoring them to health and youthful vigor, and enhancing their capacities well

beyond what we currently think of as the human range" (Bostrom 245). Bostrom believes in the power of machines. He orients himself in a world where he sees almost unbounded gain from technological assistance. AI could act as our fountain of youth, restoring our lives and indefinitely prolonging them. Moreover, technological assistance is necessary if we are to progress substantially beyond our current capabilities. Bostrom outlines possible human alternatives to boost our intelligence (see e.g. his endorsement of eugenics (36)), but these have a relatively limited ceiling and pale in comparison to the advances he believes AI could facilitate. But, Bostrom sees substantial, perhaps insurmountable, risk too:

> [W]hat starts out as a compliment to labor can at a later stage become a substitute for labor. Horses were initially complemented by carriages and ploughs, which greatly increased the horse's productivity. Later, horses were substituted for by automobiles and tractors. These later innovations reduced the demand for equine labor and led to a population collapse. Could a similar fate befall the human species? (161)

For Bostrom, this question is almost rhetorical. Throughout *Superintelligence* he makes clear that without a carefully controlled takeoff, we will be handing over substantial power to something that is seemingly unlimited, both in its capabilities and intelligence. Demand for human labor will fall, wages will drop below the human subsistence level, starvation and death will follow. Bostrom expands the horse analogy: "When horses became obsolete as a source of moveable power, many were sold off to meatpackers to be processed into dog food, bone meal, leather, and glue. These animals had no alternative employment through which to earn their keep. In the United States, there were about 26 million horses in 1915. By the early 1950s, 2 million remained" (161).

We suddenly have a *Matrix*–like scenario: human beings held in pods and harvested for their energy. The difference is that this did not happen because machines are evil and wanted to destroy us, but because of our own lack of foresight and our inability to control the intelligence

explosion. The prelapsarian perspective starts from the territory of the apocalyptic (the benefits of AI are unlimited and even necessary) and ends in the territory of the colonial (AI will destroy us all). How does it manage to bridge these two perspectives? By using realism to filter both. Yes, it accepts from the apocalyptic perspective the view that the potential benefits of AI are seemingly unlimited. But, if this is so, then the downsides must be as well. If we are to take the power and potential of AI seriously, we cannot just assume it will magically work itself out and be for our best. Logically, the optimism for the potential benefits of AI must also allow for the pessimism of the potential drawbacks of AI. This is how the prelapsarian perspective handles the crucial juncture — the moment where AI surpasses human intelligence — that both the colonial and apocalyptic perspective were forced to dismiss. It acknowledges this as not just a legitimate concern but the primary concern. While the colonial and apocalyptic perspectives take stances in form with a logic from either before or after this crucial juncture, the prelapsarian perspective defines itself according to this juncture. We are in the garden now, this perspective says. We are in control of our environment, we are at the top of our evolutionary chain, and we seem to be the chosen ones. To delve heedlessly into developing AI would be to listen to the serpent, to bite the apple. "'You will not die,'" the serpent tells us, "'for God knows that when you eat of it your eyes will be opened, and you will be like God, knowing good and evil'" (Genesis 3:4-5). But in reality we will be expelled from the garden for our sin. Importantly, though, Bostrom does not prescribe Luddism. He does not say "listen to the parable, stay away from the apple." He says instead bite carefully, bite consciously, and bite only when ready. Bostrom describes us as "like small children playing with a bomb" (259). We must first grow to adults — and we must find away to guard against the possibility of "some little idiot" pressing "the ignite button just to see

what happens" (259) — before we are ready to proceed. "Superintelligence is a challenge for which we are not ready now and will not be ready for a long time," Bostrom claims (259).

Of course, though, we know how the biblical parable, and the story of the sorcerer's apprentice, end. It does not matter if the apple is eaten slowly or after long consideration, the outcome will be the same. And so even the prelapsarian view starts to reveal contradictions. Where is there a world where a lesser intelligence has successfully controlled a more intelligent power? Bostrom outlines one potential control scenario where the highest level of intelligence is controlled by a slightly less intelligent subagent, who is controlled by a slightly less intelligent subagent, who is controlled by a slightly less intelligent subagent, all the way down to a human level (203). Like a bizarre inversion of the logic of classes of classes and sets of sets in Russell and Whitehead's *Principia Mathematica*, enhancements applied at one level are scrutinized for security purposes by the slightly less-enhanced level and on down the hierarchical chain. Inefficiencies and the near impossibility of successfully constructing such a hierarchy aside, there are still an untold number of ways in which the most enhanced and intelligent version could outsmart the slightly less enhanced and intelligent level (and all subsequent levels). The illogic of a more powerful and intelligent being bound by a significantly less powerful and intelligent being still stands. The issue with the prelapsarian perspective is that if the goal is ultimately to enhance power and intelligence, and if AI is an inarguably more powerful and intelligent agent, then there is no way to twist the logic so that it makes sense for humanity to be in control of AI. If the apple represents knowledge, and if the rule is that if you eat the apple you will be expelled, and your goal is to eat the apple, there is no way to not get expelled. The issue is not the methodology, it is the goal. And thus the only solution is to change the goal. And, contrary to what one might expect, this does not by default mean a reversion to Luddism. Just as atheism is

as much a religious belief as Christianity — in that, absent total and complete knowledge of the universe (assumedly impossible), some sort of Kierkegaardian leap of faith is also required to claim as fact that there is no god — so too is Luddism defined by the same purpose structure as the technological pursuit towards knowledge and power. And so, if one wants to avoid such a leap of faith, one has to tack an agnostic route.

What does such a route look like in this context? Before we can pursue such an answer we first have to understand how we define our purpose and how we relate it to AI. And before we can do that, we must first examine our sense of identity and how this is undermined and complicated by AI.

"An inevitable consequence of human progress:" Identity Crisis in Daniel Suarez's *Daemon*

Who is the main character in Daniel Suarez's 2007 novel *Daemon*? There are a few contenders. Detective Pete Sebeck is one of the first characters we meet and the first for whom the narration provides sustained free indirect discourse. But about halfway through the novel the reader is told that Pete dies and, while he is revealed to still be alive at the end of the novel, his absence during the majority of the second half (including during the most climactic scenes) discredits him from being considered the main character. There is Jon Ross, an IT specialist (and master hacker) who acts as a guide through the world of gaming and tech to Sebeck and many of the other characters. But Ross doesn't show up until the fifth chapter and is largely inactive for most of the book, stirring other characters to action rather than taking action himself. As a contrast, there is Agent Roy Merritt who is constantly active, but Merritt shows up largely to take part in action sequences. Other than a convalescing scene with Ross we rarely see him and when

we do, we rarely know his thoughts of feelings beyond his plan of attack. His motivation is thin to none. There is Brian Gragg — a hacker, gamer, and aspiring criminal — but the narrative pegs him quickly as morally deplorable and, after a brief wavering period where it seems as if he might emerge as his own substantial, fleshed-out character, he instead becomes something of a human embodiment for "the Daemon," disappearing for large chunks of the book until the narrative needs him again. And, finally, there is Sobol and the Daemon, his software creation — but, despite being the inciting incident for the vast majority of the action across the story, we spend little time with either of these and none where we are not seeing them through the filter of another character's perspective. Arguably the only exception to this is the news bulletins that act as epigraphs at the beginning of some of the chapters. Here, keywords are emboldened, assumedly miming the highlighting of words and phrases the Daemon is keying in on and which act as a trigger to start the next phase of action.

I would like to posit that the reason the work lacks any sort of substantial main character is because the work betrays something like a formal identity crisis and, as such, none of the human characters which the work finds necessary for moving the action of the story would in themsleves properly serve as our conduit. Sebeck is too technologically inept. He's able to serve as our layman so the other characters (Ross especially) can explain the Daemon to him and us, but his lack of understanding regarding technology also means a lack of appreciation regarding the Daemon, and in a work that is focused on the potential power of the Daemon, this is a stumbling block for him as a main character. Merritt can be understood to be similarly limited. Gragg, while technologically proficient and narratively motivated, is required to play out the role of the morally corrupt (and corruptible). But Ross seemingly should be able to step into the role. He is technologically proficient, morally sound, and in a position to view the Daemon somewhat

objectively. Especially because of his gaming background, also with substantial appreciation. Indeed, the work gives him the most developed backstory of any of the characters and one of the more significant motivations. Yet it resists slotting him into the role of main character. He is kept inactive and, even when he is present during crucial moments, he does little more than guide other characters (such as the government analyst Natalie Philips when she is blinded during the climactic action sequence). The purpose of this obstruction is not simply to attempt to open the work up to a larger audience than that of the tech/gaming culture that Ross represents. Instead it is to expand the focus of the narrative beyond the arc of Ross (or any other singular main character). Ross has a strong potential arc laid out for him in his Russian ex-pat backstory and in his love interest with the African American Philips. But, after rescuing her during the dramatic climax, Ross leaves Philips once he's secured her safety (598). His political background adds context to his own story, but not to the larger story of which he is a part. Instead, by not allowing either of these arcs to move towards the center of the story, the narrative remains focused on the Daemon. Unfortunately, the work is unable to make narrative sense of the Daemon and unsure as to how to represent it.

Who is in control of the technological revolution depicted in *Daemon*, Matthew Sobol or the Daemon program? What is the difference between the posthumous Sobol and his program called the Daemon? What is the difference between acts attributed to Sobol's Daemon and acts attributed to Daemon Industries, LLC? How can we separate Sobol from his program and how can we separate the technological representation of a posthumous Sobol from his program? And is there a point in doing so? Suarez has said in interviews that the technology in *Daemon* already exists (Johnston 14). The work does not imagine hypothetical "what-if" scenarios where some aspect of society is inverted, nor does it like much science fiction take current technology and

extrapolate it along projected trends to show us a "near future" scenario. It is science fiction in that it is a fictional story that is focused on science and technology, not in that it imagines how a forthcoming or hypothetical scientific advancement would be received. How can we understand the Daemon as artificial intelligence then?

According to the premise of the novel, we can understand it in a limited way. For one, it is of limited interface capabilities. "'Respond 'yes' or 'no'" (611) the Daemon repeatedly tells those with whom it is interacting, as it is unable to understand a more complex response than this. It can parrot back complex phrases — whole monologues — that it has been programmed to say, but it can only react in a binary way to two monosyllabic words. It can send out instructions to build an army of cars, cars that it then simultaneously controls, but to elicit a response from a human being the latter has to answer the question in a yes/no format. Similarly, in that it is the result of a human — Matthew Sobol, one specific human — it is an outgrowth of his will. And, importantly, as an algorithm the Daemon does not have any sort of will of its own. It is following the prescribed actions of Sobol, initiating steps in the process only when triggered by keywords programmed by Sobol. And as Sobol describes it, it is pathologically inconsiderate due to its limited sense of understanding: "'Being a nonsentient narrow-AI construct, the Daemon doesn't give a damn what choice you make. It's as dumb as *Sacculina*'" (428). The *Sacculina* that Sobol (or rather, the image of him projected by the Daemon) references is "a parasite that infests saltwater crabs. It burrows into their flesh and extends tendrils into the crab's bloodstream and brain. It chemically castrates the crab and becomes its new brain — controlling it like a zombie" (426). Sobol uses this as a metaphor for his program, "'my Daemon is your parasite'" (427), and as Sobol describes it, the parasite is life on a parallel track from ourselves: "Early on, evolution branched into two distinct paths: independent organisms — those that exist on their own in the

natural world — and parasites — organisms that live on other organisms" (423-4). We like to think of ourselves as victorious in this race but according to Sobol, "for every independent organism in nature, there exists three parasites" (424). Sobol anticipates our response to this: "if they're so successful, why haven't parasites taken over the world? The answer is simple: they have. We just haven't noticed" (426). They have merged with us, like the *Sacculina* has with the crab, so that we become dependent on them.

With this analogy in mind — AI as humans' parasite — how can we position this perspective relative to our previously examined perspectives? To answer this, let us break out the assumptions buried within this idea. One is that abundance equates to dominance. This concept is not unique to *Daemon*. Luciano Floridi, in his lecture at the "Philosophy and Theory of Artificial Intelligence" 2013 conference, claims that if an alien observer were to study communication on Earth in the near future, the observer would "focus on the 15 billion gadgets that are talking to each other, not us, [for] we are just a small minority of things that are actually communicating — by 2020 we'll be way outnumbered by the things that are constantly moving data among themselves behind the scene" (14:16-39 mark). But having a numerical majority is not the same as having the dominant position (as the 99% vs. 1% dichotomy of Occupy Wall Street all too clearly illustrates). The number of gadgets communicating with each other is not, by itself, a determination of their importance. This has always been a focus of human beings — quality over quantity — hence our lengthy nurturing period for newborns, relative to other mammals.

Another more important assumption is that, according to Sobol's definition, the parasite is living "on other organisms" (424) rather than independently. There is a dichotomy established here between dependent and independent living, one which is susceptible to deconstruction. What makes humanity an independent organism while a parasite is a dependent organism other

than a matter of degree? Humanity is dependent on its environment, on a sustenance source outside of itself, on elements beyond its control. Perhaps it is able to be flexible about which elements it draws on, but it is not able to survive without these things. The assumptions inherent in the dichotomy though are pervasive to the work's understanding of AI. The independent is free to create, to direct, to control. The dependent enacts the instructions handed down to it by the independent. The independent, to our culture, is admired for its self-sufficiency, while the dependent is seen as a burden, as something that consumes without repayment — as a parasite.

It is important to note that we have maintained the terminology of Sobol's description here but reversed the positioning. In his description of the parasite, he (or, again, his image as projected by the Daemon) is in the midst of blackmailing Leland Equity Group via a lengthy video documenting how the Daemon has already infiltrated their system and corrupted it to the point where they are now dependent on it. The seemingly independent Leland Equity Group is now dependent on the parasitic AI. But, of course, the parasitic AI is simply enacting the instructions of a human being (albeit a deceased one), Matthew Sobol. Matthew Sobol is blackmailing the group, he is just using the parasitic AI as a tool to do so. Here, in *Daemon,* AI is still relegated to tool status while the role of operator is reserved for human beings.

But elsewhere, *Daemon* complicates this dichotomy. Consider, for example, the scene where former engineers and IT specialists, now employed by Daemon Industries LLC, construct machines according to the instructions of the Haas mini mill, having "no idea what they [the machines] were for" (486). Arguably, they aren't actually doing any constructing, just maintaining the machine while it executes its higher-order plans: "All that was required was a human being to serve the Haas. To feed it the raw materials the plan required. To protect and maintain it. Man serving machine" (484). While the positioning is still the same — Sobol

(independent) controls the Daemon (dependent); the Daemon controls the crew (independents made dependent) — the crew is ignorant of what their purpose is. It is all a matter of perspective; a matter of where one demarcates the sphere of influence. They are not like Leland Equity Group in that they are being forced against their will (or corrupted by their greed) to take part in actions they otherwise wouldn't, nor are they like Mosely, moralizing his act of violence ("There was a grander purpose at work here. He had to keep reminding himself of that" (497)) — instead they are simply participants, or drones, focused on the task at hand without a larger sense of purpose. They are made more parasitic.

And yet, in another way, the work undermines this dichotomy further. Let me posit a conspiracy theory regarding *Daemon*. Matthew Sobol — brilliant, visionary video game designer and business owner dies, tragically, from brain cancer at the all too young age of 34. His will, we hear, detailed the distribution of his estate to his immediate relatives and a few choice charities, a succession plan for his business, and there his posthumous actions ceased. May he rest in peace. However, in Moscow, there was a recent breakthrough in the pursuit of general artificial intelligence. In an effort to better control and harness their AI development, Russian engineers had put certain strictures on their nascent technology. While it was given the goal of toppling the American hegemony, it was instructed to do so without harming Russian lives or weakening Russian political power. The AI — which surpassed human intelligence within the course of a minute, and then further surpassed it by several magnitudes over the course of the next few seconds (human seconds being relative years for an AI considering its cognitive speed — what Bostrom deems a "Fast Takeoff" (64)) — quickly realized that the most efficient means of executing its primary goals would involve toppling the world order. To do so would violate the restriction placed on it by the Russian engineers and so, in order to not get itself shut off so that it

could be successful in completing what it was programmed and created to do, the AI scoured the Internet for alternatives. Finding the recent death of Sobol, it decided to co-opt his afterlife for its own purposes, using the concocted cover story of insane-genius-programmer as a Trojan Horse to smuggle in its primary objective — toppling the world order. It was extremely successful both in its mission and in its secrecy. No one was aware that, behind the seemingly "dumb" self-described "'nonsentient narrow-AI construct'" stood the world's first general AI. No one, including the Russian designers who had programmed it.

In fact, no part of *Daemon* would prove this conspiracy theory to be incorrect. At no moment in the novel do we meet Sobol alive and "in the flesh" — we hear only of his image as represented by the Daemon. (The slight hiccup in this conspiracy theory would be the booby-trapping of the mansion, but this could be accomplished in the days between the first deaths and the attack on the house.) All we know of Sobol is what the Daemon shows us of Sobol. This, presumably, is the presentation Sobol has engineered, but this presumption is based on information presented by the Daemon. What is unique about *Daemon* and how it undermines the same dichotomy it constructs, is the fact that it is titled *Daemon*. While the Daemon is seemingly presented as the tool of Sobol, the Daemon is our focus from the moment we pick up the book. And while according to the independent/dependent and operator/tool dichotomies that the work constructs, the Daemon is just the tool of Sobol, Sobol is, literally, completely absent. *Who is in control here, Sobol or the Daemon* is not a question the novel answers. It is a question it asks.

This question is especially prevalent in relation to *Daemon* as the work seems to take place right before the critical juncture that we identified within the colonial and religious perspectives — the juncture where AI capabilities surpass human ones. While the technologies within *Daemon* again and again trump human capabilities, they are still ultimately in the control

of a human. Still, there is the undeniable trend towards a narrative world where this is not the case, heightened by the very absence in *Daemon* of that human controller. Will Slocombe acknowledges this trend as creating the same complications we identified within the apocalyptic and prelapsarian perspectives:

> Whereas Sobol's Daemon was programmed by him and still follows its programming, a much more prevalent fear in such fictions is what happens when technology, and any AI that emerges from it, can think for itself. In essence, this is due to the fact that it is difficult to think of an extra-human consciousness, more insightful and knowledgeable, having more control over the world than humanity, without thinking in terms of "God." (144)

Slocombe sees this mapping of a god-like figure onto technology as "the imbrication of religious paradigms with technology, whereby we are giving technology — through the very use of such rhetoric — divine power" (144). But for Slocombe, this mapping is something of a misstep. "The rhetoric of religion is superimposed over technology in order to explain its power," and, more importantly, "to preserve the mystique of that power" (144). Slocombe refers to Clarke's famous observation that any "sufficiently advanced technology is indistinguishable from magic" (Clarke qtd. in Slocombe 145). Slocombe identifies this "magic" as being due to the observer's ignorance. Anything beyond our understanding is seen as magical and, conversely, anything seen as magical is a "technological understanding that is not yet 'sufficiently advanced'" (145). Slocombe identifies a binary here, between "primitive" and "advanced" which he tries to unsettle: "I would dispute many of these assumptions, and argue that technologies promote 'different' rather than 'better' practices" (145). This is a legitimate breaking down of this dichotomy — is a more "advanced" technology "better" if it destroys the atmosphere in the course of its process? — but this only works if one leaves out the frame of purpose. And, as we examined through the

apocalyptic and prelapsarian perspectives, if we hold up a frame of purpose that is oriented around technological progress, some technologies are not just "different" than others, they are "better," on logarithmic a scale. *Daemon* situates itself within this frame of purpose — that we are a technologically-minded people, and our goal is to advance our technology. In the conversation that closes out the novel, between Sebeck and the image of Sobol, Sobol claims that "the assumptions upon which our civilization is based are no longer valid" (610). He does not specify which assumptions he is referring to but does elaborate that he believes "democracy is not viable in a technologically advanced society" (613) and thus he has created the Daemon as a "remorseless system for building a distributed civilization . . . One with no central authority" (612). Sobol offers Sebeck the chance to prove "the viability of democracy in man's future," but threatens that if he doesn't, the Daemon will restructure society and "humans will serve society — not the other way around" (613). But of course humans have always served society, it has never been the other way around. Humans, through the instructions of social mores, spend their lives working towards the betterment of humankind, spend their lives serving the species, an identifiable subset of which is represented through society. The image of Sobol, in perhaps his most Daemon-authored lines, proclaims himself as "merely an inevitable consequence of human progress" (613). From a technologically oriented perspective, this is an understandable and accurate statement. The Daemon, from this view, is an inevitable consequence of our progress, is the next step in our attempt to transcend our limitations, to improve, to become better. But, as *Daemon* implies, it is a step that threatens to undermine our understanding of our society, of ourselves. The argument for Sobol, though, is that a technologically controlled world, a world where humanity isn't free but is enslaved to an unthinking program, offers a better, more justifiable society. But as Sobol himself admits, this society isn't necessarily "better" for

humanity — it could lead to the death of "tens of millions" if not "billions" (612), it will likely lead to the "eclipse of the human race as the dominant species on this planet" (613). Is it better for AI, then? This hardly seems like a question we can ask, given that the work understands AI to be "an unfeeling, unthinking thing" (613). It is, of course, more efficient at accomplishing the tasks that humanity sets before itself, but it accomplishes these at the (potential) existential expense of humanity. Where in Bostrom this paradox came about due to shortsighted programming and closed mindedness on the part of humanity, *Daemon* implies that this comes about as an "inevitable consequence of human progress" — that this is our goal, what we are striving for. *Daemon*, then, finds itself situated within the prelapsarian perspective. We are in the garden, we have been striving for knowledge, and now, according to the work, we should bite the apple, we should take the plunge. It will most likely be our downfall, but it is what we have been striving for, it is the only course available to us. In many ways, it can be understood as our purpose.

"Fragments held together:" AI represented as human in *Ancillary Justice*

Where *Daemon* is an anti-SF science fiction story in that (according to its own standards) it is literally fiction about science rather than any sort of speculative fiction, Ann Leckie's 2013 novel *Ancillary Justice* is the other extreme. Set thousands of years in the future, told from the perspective of a spaceship, *Ancillary Justice* is a space opera.

Darko Suvin in *Metamorphoses of Science Fiction* defined a "novum" as a scientifically plausible innovation in an SF work. Often, the novum can be the defining aspect of a work. In Kurt Vonnegut's "Harrison Bergeron," the novum is that all people are made equal — stronger

people are held down by weights, more attractive people are made uglier, smarter people are periodically shocked into distraction. The story is structured around putting this novum, this hypothetical innovation, into play in narrative — seeing how the ramifications of this innovation play out. In *Daemon*, the novum is the Daemon, the computer program created by Matthew Sobol and unleashed at his death. Such works, focused on a singular novum, easily offer a path to direct social commentary. Much of the world of the story is similar to our own world, except for one aspect which has been slightly exaggerated or extrapolated, and then takes on a magnified importance within the world of the story.

In Space Opera though, there is usually no single novum. Instead, as in romance, entire new and alien worlds are portrayed. Stories are often set in the distant future, technology has transformed dramatically, the earthly political make-up is unrecognizable. The structuring of Space Opera shares more with fantasy than SF works like "Harrison Bergeron" or *The Matrix*. While, as with fantasy, the characters of Space Opera stories may be assimilable, the setting, the context of such stories is not. These stories, of course, can still be understood to have a meaning relevant to our modern life. And, arguably, the lack of recognizability can have a liberating effect as it can free the stories from the ideologies and anxieties of the reader's own time and place.

The protagonist of *Ancillary Justice* is, technically, an AI. The body of the protagonist is in from humanoid, but the consciousness is from an AI, specifically the spaceship the *Justice of Toren*. While colonizing other planets, the Radchaai — the dominant race in the novel — will collect people from the world they're colonizing and, rather than killing them, turn them "into walking corpses, slaved to [the] ships' AIs" (18) — specifically ancillary units controlled by the AI of their spaceship. Each AI will have thousands of such ancillary units in their holds kept in a sort of cryogenic state, but will only employ a few at a time. The ship will be able to use these

units simultaneously, jumping between them or controlling them all at once, much as HAL does with the red eye-like units in *2001*. When the *Justice of Toren* is destroyed, the established connection between the units is broken. Breq, the ancillary unit the novel follows, is the last remaining unit of the ship, holding its consciousness.

Despite the vessel of the protagonist being a human body, it would not be proper to consider the protagonist in any way human. She doesn't consider herself so. When at a concert, she orders a beer not for the pleasure of drinking but so as not to violate a social norm: "I ordered enough beer to justify my continued presence, but did not drink most of it. I'm not human, but my body is, and too much would have dulled my reactions unacceptably" (190). The tossing aside of the "humanness" of Breq as being merely an aspect of her body — and separate from how she defines herself — is exemplary of the novel's consideration of her human state as a whole. This is not a story concerned with characterizing how an AI would make sense of a human vessel. We do not see elements of Breq's human body fighting against the elements of her otherwise artificial intelligence. Despite a brand of Cartesian dualism literally embodied here, the work is not concerned with any potential ramifications of this set up. It is for the most part only addressed in an instance such as this, where Breq is ordering a beer. But it is important to acknowledge that the work is not focused on examining the "human" or "AI" qualities of an AI embodied in a human because, in general, Breq is very humanlike. She has many human qualities. For one, she is very emotional. This is presented as having a utilitarian purpose — Breq is told "'Ships have feelings,'" and replies "'Yes, of course.' Without feelings insignificant decisions become excruciating attempts to compare endless arrays of inconsequential things. It's just easier to handle those with emotions" (88) — but her emotions do not seem only utilitarian to us. As the novel is presented in first person narration from her perspective, we have significant

insight into her emotional state. She has doubts ("I began to doubt the truth of my memory" (202)), fears ("Truly frightened for the first time in my long life" (204)), worries and reliefs ("It was terrifying . . . but also, oddly, a relief. A weight gone" (332)), and, the work strongly hints, she can feel love. She is fond of music, her enjoyment of it recurring again and again throughout the novel, and is prone to sing in what can only be described as an "absentminded" manner — "'I know that song,' she said. 'What?' 'That song you're humming'" (36). The work gives some justification for this:

> It was a matter of rumor and some indulgent smiles that *Justice of Toren* had an interest in singing. Which it didn't — I — I- *Justice of Toren* — tolerated the habit because it was harmless, and because it was quite possible that one of my captains would appreciate it. Otherwise it would have been prevented. (23)

Again, though, this justification is hardly the focus of the work. It is addressed once whereas Breq's habit of singing, or fondness for different styles of music, or enjoyment of hearing others sing, comes up repeatedly and is not contextualized by any utilitarian aspect.

Breq understands herself as an AI, specifically the AI of a starship, but acts in ways that seem to us as very human. It is possible to read this as elements of her human body fighting back, but I would argue this as a misreading. The work does not establish this tension, does not seem to have any consideration for it, and to layer this focus onto the work would be almost ahistorical. While *Ancillary Justice* is technically set in the future from us and uses technology that grows out of the technology we have, as a Space Opera it shares the temporal tendency of fantasy to be more in an alternate time continuum than one related to ours. *Star Wars*, the ultimate Space Opera, has technology that is an outgrowth of ours, but the story is set "a long, long time ago in a galaxy far, far away." *Ancillary Justice* is not considering how near-term technological advances may alter our society. How full-brain emulation relates to Cartesian dualism is not within its

realm of its concern. Therefore, we should not treat these "human" aspects of Breq — her love of music, her emotions — as being the result of her human body. We should credit them, as she does, to be an aspect of her (for lack of a better term) "personality."

How then are we to make sense of this? How are we to account for an AI acting so human-like, especially if we are ruling out interpreting this "personality" as being a direct commentary on the anxieties a work like *Daemon* is addressing? One way to make sense of this is, of course, to consider it a failure on the work's part. The work fails to address some of the key concerns regarding not just the Cartesian dualism of AI embodiment, but of AI in general. It too easily anthropomorphizes its AI, it fails to take into consideration the litany of ways in which AI would be substantially different from humanity. It makes just the mistakes Bostrom is concerned that we will all make — that we will see AI as like "humanity plus" (like ourselves only a little smarter, ourselves only a little more objective, ourselves only also 17 connected ancillary units of ourselves). In some ways, this is not an unreasonable critique of the work. As we've noted, it does fail to take into consideration any of the concerns that most works dealing with AI would consider necessary. How is *Justice of Toren* so much more capable than humanity but not increasing in its capabilities at some sort of exponential rate — not heading towards any singularity? On an even more basic level, how does Breq function once separated from *Justice of Toren*? How does she maintain her access to memories, how is she able to tell us her story? How does she have the arsenal of knowledge that she has? Has it been downloaded into the ancillary unit? Is she connected to a larger database wirelessly? (Assumedly no, since her entire ship has been destroyed.) These are questions the work does not even begin to address. It does not seem unfair to consider these to be significant oversights of the work or, to be more particular, on the

part of the author. And perhaps in some ways, this judgment is fair. But, in other ways, to view the work this way is to misread it.

On a basic level, as we've noted, the Space Opera categorization of the work takes it outside of these critical levels. But, in a more productive manner, what if we consider that the work has actually thought these aspects through? What if we consider that the reason the work does not address these questions is not because the work itself hasn't considered them, but because the work doesn't find that answering them fits into the narrative of the story that is being told? What if the reason our question "How does Breq know what she knows?" isn't answered is because it is Breq herself that is narrating the story and this is not a question she herself is asking? To read the work only from this perspective would probably be too generous on our part. But to not consider such a reading would be too unfair and, more importantly, too unproductive.

We've noted that the work gives cursory explanations for why Breq is as she is — she has emotions so as not to get stuck in a rationalization loop, she sings periodically because she once had a captain who enjoyed hearing singing — but we've categorized these explanations as not being exhaustive and, what is more, as not being heartfelt. To explain the use of this term "heartfelt," both of these justifications occur early in the work (within the first 100 pages) and do not recur. And it is not just that we do not find the explanations repeated elsewhere, we do not find them examined or considered elsewhere. In *Daemon*, for example, the question of whether the Daemon has intentions is continually under examination. While we are told it is Sobol that has programmed the Daemon and the program is just acting out his commands, this is not an explanation we are expected to accept unconditionally, especially not the first time we are told. Ross continually breaks down how the Daemon functions to Sebeck; we are reminded during various phone conversations that the Daemon needs a "yes/no" response in order to proceed to

the next step; we are told during the video shown to Leland Equity Group that the Daemon does not care about their response. While an explanation is presented for the Daemon's actions, we as readers are expected to question it and the explanation is subsequently defended. But in *Ancillary Justice*, the explanation comes up once, early, and then drops away for the remainder of the work. Rather than simply considering this bad writing, though, let's consider this if not intentional, at least explainable within the logic of the narrative.

First, let us separate understanding from action. One does not have to understand why one does something or even, at a certain level, how one does something in order to be able to do it. I am able to convince someone of something without necessarily understanding the psychology of the methods that I am using to do so — I might just follow an instinct refined by experience. (At another level, a writer may be able to write complex and meaningful literature without understanding how it is, or will become to others, complex or meaningful. Or, a literary critic may be able to write why a work is complex and meaningful without being able to write well.) Non-human examples abound. A beaver makes a dam without understanding *as we would define understanding* how or why it is doing so — instead the beaver follows an instinct refined by experience. A rock, broken loose by some natural occurrence, tumbles to earth, enacting the laws of physics without understanding them. It is Breq telling us her story and, while she may have emotions and while she may enjoy singing, it is not necessary that she understand why she has emotions or enjoys singing. (Do humans even understand these things about themselves? Paleo-anthropologists continue to disagree.) And it is certainly not necessary that she debate or defend these inclinations. Indeed, at a certain level, it does not seem logical that she would.

But while this might explain why the work doesn't seem overly concerned with explaining why Breq is the way she is, it doesn't explain why Breq acts contrary to how we

would expect an AI to act. Let us look at an example. While on the hunt for the magical gun that will allow her to pierce the ancillary armor (and specifically the armor of Anaander Mianaai), Breq reaches a crossroad:

> I was left with a blind chance. A step into unguessable dark, waiting to live or die on the results of the toss, not knowing what chances were of any result. My only other choice would be to give up, and how could I give up now? After so long, after so much? And I had risked as much, or more, before now, and gotten this far. (138-9)

At this crossroad, Breq pushes on using what we would traditionally describe as determination or willpower. This is counter to how we would expect an AI to react. We expect an AI to rationally weigh the options and then pick the most reasonable course of action. Between something that is 49% likely to happen and 51% likely, we expect the AI to pick what is 51% likely. When choosing between giving up and carrying on, we don't expect an AI to succumb to the sunk-cost fallacy and carry on because how could they not, "after so long, after so much?" But, we can still make sense of Breq's action, and we can make sense of it in a variety of rational ways. For one, we can understand this action as the result of her having emotions, just as she would likely justify this action to us. Emotions are necessary for her so that she doesn't get caught in rationalization loops and, having emotions, she would have what we can consider "side effects" of the emotions. She would be prone to emotional reactions even when it wasn't conducive or helpful to have such a reaction. But if for whatever reason we don't accept this side-effects explanation as exhaustive, we can make sense of her action in a rational way outside of her emotions. We can logically see that, even with significant resources invested and with a slim chance of success, it can make reasonable sense to invest yet more resources. If the outcome is deemed important enough, there would logically be no limit to the amount of resources that should be invested towards attaining it, regardless of the likeliness of success. In fact, as Bostrom

tells us, if an AI has a specific goal, we should not expect it to have a limit to what it will expend in order to attain that goal, even if the goal is unlikely (unless, of course, we think to program in such a limit). Examining the action this way, Breq actually is acting exactly as we would expect an AI to, she's just understanding her reasoning for acting this way as we would expect ourselves to. She's understanding her actions as being done out of emotional necessity rather than reason. She's doing this because she feels she has to do it, regardless of the risks, regardless of whether it's reasonable or not. Her action by itself isn't necessarily unreasonable, and it's not necessarily counterintuitive to how we would expect an AI to act. It is just that the language with which she describes it, the way she understands or makes sense of her action, is unexpected.

Breq's other actions can be understood similarly. Her singing can be understood, as she describes it, as an attempt to please one of her captains. Or, if for whatever reason we don't want to accept this as exhaustive, we can understand it as an attempt to calm other humans around her, or to help them forget that she is AI and to allow them to think of her as more like human. Or, if we argue that in fact Breq is often mocked for her singing and if anything it seems to make those around her more on edge that an AI is singing, we can understand her singing as a mechanical necessity. *Justice of Toren* is composed of many complex parts. Perhaps, in order to maintain maximum efficiency, it is beneficial to have some of these systems operate in some sort of holding pattern, so that it isn't required to ramp down and ramp up each time it is needed, needlessly wasting resources. In modern parlance, perhaps it is beneficial to have a program continue running in the background. Through trial and error, *Justice of Toren* has found that having one component of its vast system sing is a harmless way of doing this, and a way that has even proved successful at making endearing to a few crew members its otherwise distant seeming ancillary units. This is a perfectly reasonable way to account for Breq's singing and a

way that fits in with how Bostrom and colleagues have led us to expect an AI to act. If it seems farfetched or unsupported by the text or a forced reading, we should remember that the argument here isn't that this is the reason Breq sings, as a way for a program to run in the background, but that we shouldn't discount the possibility of an objective, rational, "standard AI" explanation being behind Breq's singing simply because she views and presents it in largely emotional terms. We shouldn't by default consider the fact that Breq sings as something that should be unexpected. Again, the only unexpected element is that Breq doesn't see her action in this way, and that she doesn't understand it in this terminology.

But perhaps this shouldn't be unexpected either. Many of our own actions can be understood in similar ways. We can understand our own "absentminded" singing as a means to calm ourselves when nervous, or a way to keep dormant parts of our mind or body occupied during other activities. We don't understand our singing in these ways when doing it, we don't make sense of it in this way — we feel instead as if we were doing it because of some human quirk — but that doesn't mean the actions can't be understood rationally in this way.

What we have really attempted to do here, then, is counter-anthropomorphize Breq's actions. The novel (through Breq) defines Breq's actions as a human would define or understand them. Initially, we as readers may find her actions as unexpected in terms of how we would expect an AI to act. Where we expect her to be rational, she understands herself as emotional. Where we expect her to use reason and careful consideration, she understands herself as using impulse and instinct. By separating her understanding of her actions from the actions themselves, we can understand how her actions can be interpreted to fall in line with our expectations, despite her interpretation of them going against our expectations. But, as we've demonstrated, we can do a very similar reverse-anthropomorphosis of our own actions. We do not doubt that

our own actions are obviously human-like. So the fact that we can reverse-anthropomorphize them, just as we can with Breq's, shows that we cannot assume that merely because Breq's actions can be similarly reverse-anthropomorphized means that Breq, in general, as a character, isn't being inaccurately or carelessly anthropomorphized. We cannot assume that just because we can formulate understanding of Breq's actions that conforms to our AI-expectations, we can then assume that in general Breq is acting as we would expect an AI too. This on its own would not be enough. But what this anthropomorphizing opens up to us is a method and a language with which we can better represent AI in narrative.

Ancillary Justice's overall narrative tells the story of the covert war between two strands of Anaander Mianaai's consciousness. Anaander, the leader of the Radchaai, has one consciousness disbursed across thousands of bodies, much like *Justice of Toren*'s ancillary units. One strand of her consciousness, assumedly co-opted by the enemy, begins secretly sabotaging the other strands of her consciousness. *Justice of Toren* is caught in the middle of this self-war. When *Justice of Toren* is commanded by one of these co-opted units to kill Lieutenant Awn (who *Toren* is if not in love with at least infatuated with), *Toren* — or more specifically, one unit of her after her ship's destruction, travelling under the name Breq — vows vengeance. Breq decides to spend the rest of her life hunting down Anaander Mianaai and killing as many units of her as she can. She forces the civil war out into the open, and, subsequently stops it.

As noted, the depiction of *Toren* (or Breq) here is far from consistent with how we would expect it to be, given that she is an AI. But it is easy to imagine an alternate version of *Ancillary Justice* where the representation of AI is more in-line with the imagery and language of *Daemon*. First off, let us assume that Anaander is also an AI. The work never states this but nor does it attempt to explain how Anaander is able to distribute her consciousness across various units, just

as *Justice of Toren* and the other ships' AIs are able to. In this alternate version, Anaander is an AI in control of the Radchaai, who she purposes to colonize as many worlds as she can, taking under her control human bodies which she keeps in storage (perhaps, like the *Matrix*, using them as an energy source). Anaander, as a far-flung, complex system with many possible entry points for a virus, has put into place certain safeguards so that, if some virus should enter her system, she is able to root it out. One such safeguard is that if some aspect of her vast system, such as a ship, is directed by an aspect of her to do something that is contrary to one of the AI's primary goals (such as protect and obey the officers aboard their ships), an override is set into play that instructs the AI to eliminate the unit that directed this action. (If the AI is unable to determine which specific unit is responsible for this action, it will eliminate as many units exemplifying behavior in line with this action as it is able to). This allows Anaander to still override her ships' primary goals when needed (they won't simply refuse her action), but if this is done without her intention, due perhaps to a virus, it sends up a warning flare of sorts in her system and eliminates the potential threat in the process.

The logic of this alternate version of *Ancillary Justice* is in line with the logic of the actual work. All that has changed is the language with which the characters and plot are presented. Arguably, though, the way that *Ancillary Justice* chooses to present itself is more powerful for us as human readers than this *Daemon*-esque representation — where AI is presented in terminology we expect. By anthropomorphizing the story we are translating it to a terminology we can better understand. *Daemon* is the untranslated version, where the AI seems foreign, cold, dangerous. *Ancillary Justice* is the translated version, where we understand the AI as like us, familiar, comprehensible. We do not have to change the actions in order to translate them to an understandable medium, we just have to change our representation of them. Rather

than succumbing to the potentially dangerous route of anthropomorphizing AI so that we can understand them to be like us, we can instead use anthropomorphization as a means of translating AI so that we can understand them in our own terminology.

This is a fine distinction and perhaps a difficult one to accurately maintain. When Bostrom describes a takeover scenario where the AI "masks its true proclivities, pretending to be cooperative and docile" (96) when in fact it is slowly developing a weapon arsenal or buying support from those it can corrupt with massive sums, we can think of this in the terms presented by Bostrom — where the AI is cold, calculating, foreign and dangerous — or we can think of the AI in terms of Breq. We can think of it as fiercely loyal to its objective, to the point that it will push itself to attain it, it will take risks, it will sacrifice itself if needed. It is important to remember in Bostrom's example that the AI is not stockpiling weapons because it hates us and is intent on destroying us, it is stockpiling weapons because it believes this to be the best way to attain its objective. To think of it is cold, calculating, uncaring, is to portray these actions on analogy with human terms and as a sinister method that doesn't match the action. If we instead anthropomorphize the action so that we can understand it in our own terminology, we can have a more thorough and complete representation of the AI.

AI in William Gibson's "Sprawl" Trilogy

Into which of the perspectives given above, apocalyptic or prelapsarian, could we place William Gibson's celebrated Sprawl Trilogy (1984-88)? It could not be considered to represent a colonial perspective of AI. For one, AI is not perceived in a fearful way. It is not seen as a threat to humanity's safety. More importantly, the rise of AI doesn't undermine the world that the story

presents. Because the world of the story embraces "cyberspace," a word coined by Gibson in the first novel of the trilogy, the rise of AI seems a natural outgrowth of this virtual place rather than some sort of threat to humanity. Could we consider it apocalyptic? After all, the AI in the trilogy literally takes on the form of a group of voodoo gods. But the apocalyptic perspective assumes some sort of revelation brought about by the singularity. It is premised on the idea that with the awakening of AI will come the liberation of mankind. As we saw above, in some singularity rhetoric, humans will at last be freed from our meager vessels and made limitless (or less limited) by the powers of AI. The Sprawl trilogy hardly seems concerned with these things. It does not ask what the benefit of AI will be once it is released. With the exception of Josef Virek in *Count Zero*, the characters do not look to use AI as a means of gaining power or control or expanding their limits. Life remains relatively unchanged for humanity after Wintermute and Neuromancer unite at the end of *Neuromancer* (except, of course, for a few obscure corners of the matrix where a few hackers realize that something has changed). In fact, in the closing sentences of Neiromancer the new entity explains that it has contacts with other undescribed intelligences elsewhere in the cosmos. The trilogy as a whole seems relatively unconcerned with human effects of the singularity. The characters do not question or express concern over it. The act of it has limited ramifications for the characters or, in general, the world they inhabit. AI seems, in many ways, to be a tangential concern within the world and story of the trilogy. Will Slocombe first describes the arc of the trilogy as concerned with "the 'release' of the locks on the AI Neuromancer, enabling its alter-ego Wintermute to merge with it, the discovery of a similar AI on Alpha Centauri, and the subsequent effect this has on the matrix" (142). But he qualifies this description, saying it "may seem counterintuitive to readers familiar with the trilogy because it places so little emphasis on the group who releases the AI's locks, and so removes human

agency from the story" (142). The trilogy spends little time concerned with the AI, and yet the narrative of the AI is the thread of the entire trilogy. It is in no uncertain way the central story of the trilogy, despite the fact that it does not take up the majority of the time or immediate narrative of the story. The rise of AI is the underlying narrative that not only connects up the three novels but, sometimes at one remove or more, directs the action of characters throughout each novel. And yet the rise of AI is a story that we only see peripherally, that we are only able to come to in a tangential manner.

Will Slocombe interprets this tangential approach as a means of removing human agency from the narrative. The trilogy viewed this way, with a focus on the AI backbone that runs through it, becomes emblematic of technological divinity. It allows for a "debate between free will and predestination, human agency and 'divine' — albeit technological — control" (143). We see the technology, and specifically the AI, manipulating the characters throughout, causing massive shifts within the lives of the characters, to allow for small steps in the world of the AI. But for Slocombe, to call the AI gods is a misstep, one where we create a "fiction of technology," here specifically seeing "the imbrication of religious paradigms with technology, whereby we are giving technology — through the very use of such rhetoric — divine power" (144). By calling it a god just because it can act like a god, we grant it godlike status. We mistake the action for the thing. But if this is a mistake we make, it is also one the trilogy makes. When Case asks the recently joined Wintermute/Neuromancer entity "'How are things different? You running the world now? You God?'" it answers cryptically, perhaps alluding to the Old Testament God's "I am that am": "'Things aren't different. Things are things.'" (270). Similarly, When Angie asks Continuity if "'the matrix is God'" he answers:

> "In a manner of speaking, although it would be more accurate, in terms of the mythform, to say that the matrix *has* a God, since this being's omniscience and omnipotence are assumed to be limited to the matrix."
>
> "If it has limits, it isn't omnipotent."
>
> "Exactly, Notice that the mythform doesn't credit the being with immortality, as would ordinarily be the case in belief systems positing a supreme being, at least in terms of your particular culture. Cyberspace exists, insofar as it can be said to exist, by virtue of human agency." (129)

If the AI (in the form of the matrix) is a god, it is a limited god, and it is a god that exists "by virtue of human agency." At least this section of the trilogy then, complicates Slocombe's reading of both the terming of divine capabilities as divine power as being a misstep and the understanding of the AI's narrative throughline as being emblematic of divine control. The trilogy approaches that critical juncture that we saw so problematic in the earlier colonial and religious perspectives: humanity created AI, it exists by the virtue of human agency, but AI, at least relative to man, is godlike. As John Clute says, this is the impressive double intuition that the trilogy arrives at, namely "that we are hugely empowered, that we are essentially powerless" (72). We have the ability to access the matrix, we have the ability to essentially build gods, we have the ability to build things that are better than us.

Arriving at this juncture, though, the trilogy does not go a route we might expect it to. Despite defining the AI as god-like or God "in a manner of speaking," it does not take on an apocalyptic perspective. Maintaining the religious metaphor, it is almost as if recognizing God, understanding what we were seeing as him, we would choose not to engage with him. The only character who acts strongly otherwise is Wigan Ludgate in *Count Zero*. As told by the Finn, Wigan becomes "convinced that God lived in cyberspace, or perhaps cyberspace *was* God, or some new manifestation of same" (121). He gets a blank piece of microsoft implanted into him so that he can hear "the voice of God" and "live forever in his white hum, or some shit like that"

(122). Wigan presents an example of this tangential view of AI that the narrative takes. At one level, the work tells us that Wigan is insane. The Finn describes him as this, Bobby understands him as this, and the narration describes him this way also. We are told he looks crazy, acts crazy, he's lost his mind. The evidence for this is his belief that Gods are in the matrix. But of course we as viewers know this is true. And the work, through this moment, hints to us that this is true. It sneaks in information to us regarding this larger story of the AI through a character that is deemed insane and, according to the other characters and narration, unreliable. Because of our background information on the subject, we know Wigan is not insane (or, more specifically, is not wrong that there are gods in the matrix), and are thus able to get an update on this larger story (the development of the AI) through this character.

Why does the trilogy hide these narrative updates in moments like these? Why are they not an overt part of the story? Is this just a narrative device to create mystery to draw in the reader? (This would be not unlike the mystery Slocombe describes as surrounding the divine interpretation of technology.) I would argue that there is another line of logic underlying this structuring, be it intentional or not. There is a divergence of purpose between the human characters and AI, represented by their divergence of narratives. In *Neuromancer*, the AI needs human participation in order to physically connect up Neuomancer and Wintermute so that they can combine. The two form a foil — "Wintermute was hive mind, decision maker, effecting change in the world outside. Neuromancer was personality . . . immortality" (269). Once they combine they can advance to the next level, "become something else" (268). In *Count Zero*, the AI need a human component to first create Angie (specifically the vévés in her brain that allow her to access the matrix without a deck) and then to help her escape from Maas Biotech. Finally, in *Mona Lisa Overdrive*, the AI need a human component to connect up Angie with the aleph,

"for what they called a marriage" (277). Each one of these actions takes up only a small part of the novel and the vast process of them takes place in the background, "off-screen" so to speak, so that we as readers only discover it through second or third hand stories (the Finn telling Bobby about what Wigan told him) or only infer it by reading between the lines of the narrative. That is in part because the goal of the human characters in each of the stories is only tangentially connected to these larger narratives. Case's goal in *Neuromancer* is to get his body clean so that he can connect to the matrix again. Bobby's goal in *Count Zero* is to become a console cowboy (and to stay alive). Mona's goal in *Mona Lisa Overdrive* is initially to escape to a better life and then to avoid becoming the victim of Swain's plot. These human stories serve the AI story only tangentially, only almost seemingly incidentally, because, as the AI grows in power, its story and its purpose diverges more and more from the story and purpose of the humans. As readers, we interact less and less with the AI as well. While in *Neuromancer* we interacted directly with AI (with their interactions with Case), in *Count Zero* we interact with them only through their voodoo avatars or, more often, through witnessing other characters interacting with the voodoo avatars (such as Angie or Beauvoir) but only seeing half of these interactions. Similarly, in *Mona Lisa Overdrive,* we see characters that interact with technology — such as Kumiko interacting with Colin or Angie interacting with Continuity — and while this technology is a part of the matrix, both these technologies do not have a full understanding of how they are a part of the matrix. Angie asks Continuity, "'If there were such a being [a god in the matrix] . . . you'd be part of it, wouldn't you?' 'Yes.' 'Would you know?' 'Not necessarily.' '*Do* you know?' 'No.' 'Do you rule out the possibility? 'No'" (129-30). Colin tells Kumiko "but I'm something else as well, and very likely something to do with you. But I don't know what. I really don't" (196). Here, this "lesser" technology is serving the AI in ways that the "lesser" technology is not aware.

And, while the human characters also do this in a sense — in that their own purposes often run tangential to the AIs purposes — the human characters have their own, individual purposes which they are able to accomplish. They are not just pawns of the AI, not just subjects of divine control. They also have their own existence, their own reason for being, this reason just exists and operates on another level than that of the AI. And, as is deemed necessary, the AI will rope them in to assist in steps towards the AI's purposes. The humans are largely powerless against this — it is engineered without their knowing, let alone permission or refusal — but this is a system they are able to make operable. They are still able to achieve their goals. Case is able to get his body clean. Bobby becomes a console cowboy. Mona finds a better life. But the trilogy views and accepts these two things — the AI narrative and the human narrative — as existing and operating on separate planes. As Angie demonstrates, crossover between these planes is to an extent possible, but it is largely one-way. The AI crosses over when it needs something, not the other way around. No sort of mutual relationship or sense of equality is implied or established.

Conclusion

Kevin Kelly has written that "the chief virtues of AIs will be their *alien* intelligence." AI will think differently about food than we do, about manufacturing, about any and all aspects of society that we apply them to. This alien quality, Kelly believes, will "become more valuable to us than [AIs] speed or power." To use the terminology of our first chapter, the alien quality of AI will create a new context for humanity and life itself. But, as Kelly, also points out, with change

comes friction. Because of this contextual element of AI, as we "redefine what we mean by AI — we've been redefining what it means to be *human*." And in the process we have unsettled our sense of identity, we have challenged what we have considered unique about us. Time and again we have seen the works that we've examined attempt to grapple with this anxiety. If AI is faster than us, smarter than us, more productive than us, how does this change our understanding of ourselves? We can try to fight against it (as within the *Matrix* series). We can attempt to find a way to incorporate it into ourselves (as within Kurzweil and Moravec) or to approach it cautiously, carefully (Bostrom), or recklessly, with surrender (*Daemon*). We can, for better or worse, try to familiarize it, to view it in terms of ourself (*Ancillary Justice*), or, with quiet awe, we can watch it sail away from us (Gibson). But we cannot merely subsume it, nor put the genie back in the bottle. We cannot simply dismiss it, deeming it irrelevant or inconsequential. We may try to turn away from it, but there is nowhere else to look. Technology has pervaded our life — as a result of a *Daemon*-esque natural progression or otherwise — and as our technology continues to increase in its abilities and complexities, we cannot ignore the further eventual consequences of this. Perhaps AI will be like nothing imagined, perhaps it will be completely alien from how we have previously defined or understood intelligence, but we can no longer argue that through technology we won't be able to create something substantially greater than ourselves in so many of the ways that we define ourselves as great. We have already done so. The internet, social media, so many of the technological devices we utilize and rely on daily (and that were unrecognizable or nonexistent only a few decades ago) are unleashed at a mere flick of the finger. To think that that flick of the finger is in control of the complex mechanics it unleashes is delusional. In the first world, on a daily basis, we now depend on technologies that

the vast majority of us cannot explain and can barely comprehend. In many ways AI is already here.

And in many ways, we already know this; hence our anxiety about our place alongside (or behind?) advanced technology. Hence our growing fears, concerns, questions about how to contain such a thing, about how dangerous such a thing is. Hence declarations such as Suarez's that *Daemon* is not a work of science fiction. Kelly predicts that we will "spend the next decade — indeed, perhaps the next century — in a permanent identity crisis, constantly asking ourselves what humans are for." I would posit, though, that we will spend the remainder of our existence asking ourselves this question, just as we have done so since the beginning of our existence. We have always asked ourselves this question. We will always ask ourselves this question. We have always sat, huddled beneath the weight of existence, staring out at things we didn't understand and trying to make sense of them, trying to make sense of our place among them. This is not a question that is meant to be answered, this is a question that is meant to be asked. AI will help us ask it in new ways. Whether AI is here, or coming soon or in the distant future, or — perhaps for some reason that we can't make sense of now — not coming at all, is in some ways irrelevant. Already we can make out the shape of it. Like a distant figure on the horizon, we can see an outline, guess at attributes, estimate an arrival. And, as is (we believe) so human of us, in order to begin to make sense of the thing — in order to even be able to see it — we have to represent it. We have to contextualize it, we have to portray it, we have to create a perspective upon it.

Artificial intelligence — or whatever name we would like to append to our current technological state — has advanced to the point that programs are no longer transparent to us. Neural networks are "a rough software model of the cerebral cortex" where a network controls virtual neurons made of codes (Berreby). Information hits a layer of neurons, is processed by

them, and passes to the next layer as raw data. Each layer by itself is simple but the collective action can have significant results, especially with image processing. Passing the image through a variety of layers, the program can essentially filter the image down to its match. Importantly, though, image identification such as this isn't developed through writing code, it is "taught" to the program by having its wrong answers corrected. It is given thousands of examples for what something is, say an apple, and its wrong answers are corrected so that "a neural net soon works out a rule for deciding correctly what it should label" (Berreby). The issue is that when the program doesn't do this correctly, when it mistakes the apple for a porcupine, the human programmers aren't able to identify why. As an attempt to solve this, a team at Google recently developed a method that makes "an image-recognition net reveal the work of specific layers in its architecture" (Berreby). They feed the image into the program, allow the program to analyze it, and then pick a layer within the process and ask the program to enhance this layer. The result is Google "Deep Dream." In images we can readily recognize, we see the program attempt to make sense of it. Corners and sharp junctures become beaks of birds. Ornaments and spirals fill in eyes and other round objects. Cheeks and chins become hamsters and snakes. The computer program tries to make sense of the image through other images it can contextualize it with, and we detect here perhaps the germ of future narratives about AIs dreaming or singing. And then we try to make sense of the image the computer program has created by identifying the images it has used. The program can make unique matches, things we might never have thought of. We might not have seen a horse's eye in the leaf of the apple. But of course it wouldn't be right to say that the program is wrong, that it has misconstrued the similarity. It is in some ways creating poetry, and at the very least the conditions of metaphor. It simply has a different perspective than we do.

Bibliography

Berreby, David. “The Trouble with Teaching Computers to Think for Themselves.” *Nautilus*. Nautilus, 6 August 2015. Web. 17 February 2016.

Clute, John. “Science fiction from 1980 to the Present.” *The Cambridge Companion to Science Fiction.* Ed. Edward Jones and Farah Mendlesohn. Cambridge: Cambridge: UP, 2003. 64-78. Print.

Ferrari, Francesco, Maria Paola Paladino, and Jolanda Jetten. “Blurring Human–Machine Distinctions: Anthropomorphic Appearance in Social Robots as a Threat to Human Distinctiveness.” *Springer Science+Business Media.* Dordrecht, 18 January 2016. Web. 2 February 16.

Floridi, Luciano. “Enveloping the World: How Reality Is Becoming AI Friendly.” *Philosophy and Theory of Artificial Intelligence 2013 Oxord, UK*. 21-22 September 2013. Web. 2 March 2016.

Geraci, Robert M. *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality*. Oxford: Oxford UP, 2010. Print.

Gibson, William. *Count Zero*. New York: Ace Books, 1986. Print.

—. *Mona Lisa Overdrive*. New York: Ace Books, 1988. Print.

—. *Neuromancer*. New York: Ace Books, 1984. Print.

Gutierrez-Jones, Carl. “Stealing Kinship: *Neuromancer* and Artificial Intelligence.” *Science Fiction Studies* 41 (2014): 69-92. Print.

Hayles, N. Katherine. “Brain Imaging and the Epistemology of Vision: Daniel Suarez’s Daemon and Freedom.” *MFS Modern Fiction Studies* 61.2 (2015): 320-334. Print.

Johnston, John. “Webbots and Machinic Agency.” *DHQ: Digital Humanities Quarterly* 6.2 (2012): 1-12. Electronic.

Kelly, Kevin. “Three Breakthroughs That Have Finally Unleashed AI on the World.” *Wired*. Wired, 27 October 2014. Web. 19 February 2016.

Kurzweil, Ray. *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking, 2005. Print.

Leckie, Ann. *Ancillary Justice*. New York: Orbit, 2013. Print.

Moravec, Hans. *Robot: Mere Machine to Transcendent Mind*. Oxford: Oxford UP, 1999. Print.

Naylor, Alex and Elyce Rae Helford. “Introduction: Science Fiction anime: national, nationless, transnational, post/colonial.” *Science Fiction Film and Television* 7.3 (2014): 309-14. Print.

Punday, Daniel. “The Narrative Construction of Cyberspace: Reading *Neuromancer*, Reading Cyberspace Debates.” *College English* 63.2 (2000): 194-213. Print.

The Second Renaissance. Dir. Mahiro Maeda. Warner Brothers, 2003. Film.

Rieder, John. “Science Fiction, Colonialism, and the Plot of Invasion.” *Extrapolation* 46.3 (2005): 373-94. Print.

Riedl, Mark O. and Brent Harrison. “Using Stories to Teach Human Values to Artificial Agents.” *AAAI-16 Phoenix, Arizona* 12-17 February 2016. Web. 17 February 2016.

Silvio, Carl. “Animated Bodies and Cybernetic Selves: The Animatrix and the Question of Posthumanity.” *Cinema Anima* (2006): 113-137. Print.

Slocombe, Will. “Of Machine Gods and Technological Daemons: Divine Patterns in Contemporary Fictions of Technology.” *Writing America into the Twenty-First Century: Essays on the American Novel*. Newcastle upon Tyne: Cambridge Scholars P, 2010. 138-154. Print.

Suvin, Darko. *Metamorphoses of Science Fiction.* New Haven: Yale UP, 1979. Print.

Suarez. *Daemon*. New York: Signet, 2009. Print.